AMERICAN PHILOSOPHICAL QUARTERLY

Edited by
NICHOLAS RESCHER

With the advice and assistance of the Board of Editorial Consultants:

VOLUME 27/NUMBER 3 JULY, 1990

CONTENTS

AMERICAN PHILOSOPHICAL QUARTERLY, 1990
PUBLISHED BY NORTH AMERICAN PHILOSOPHICAL PUBLICATIONS, INC.
IN COOPERATION WITH THE PHILOSOPHY DOCUMENTATION CENTER
ISSN 0003-0481

AMERICAN PHILOSOPHICAL QUARTERLY

FOUNDED IN 1964

POLICY

The *American Philosophical Quarterly* welcomes articles in English by philosophers of any country on any aspect of philosophy. However, only self-sufficient articles will be published, and not news items, book reviews, critical notices, or "discussion notes" (short or long).

MANUSCRIPTS

Contributions may be as short as 2,000 words or as long as 7,000. All manuscripts should be typewritten with wide margins, and at least double spacing between lines. Footnotes should be used sparingly and should be placed at the end of the paper, numbered consecutively. They should also be typed with wide margins and double spacing. Submissions should always be made *in duplicate*. Only papers whose authors certify that, while under consideration with us, they will not be submitted elsewhere can be considered.

COMMUNICATIONS

Articles for publication and all other editorial communications and enquiries should be addressed to: The Editor, *American Philosophical Quarterly*, Department of Philosophy, University of Pittsburgh, Pittsburgh, PA 15260. Other correspondence not dealing with subscriptions should be addressed to Ms. Dorothy Henle, Operations Manager, *American Philosophical Quarterly*, c/o Department of Philosophy, University of Pittsburgh, Pittsburgh, PA 15260. (For subscriptions, see below.)

REPRINTING

The *American Philosophical Quarterly* regrets that it cannot make reprints available to authors. However, authors have the journal's permission to reproduce limited numbers of their contributions for the use of colleagues and students. (Two copies of the relevant issue will be provided gratis for this purpose.)

While the journal holds the copyright on materials published in its pages, it routinely accords contributors permission to reprint in books or anthologies authored or co-authored by themselves. All other requests to reprint should be addressed to the editor.

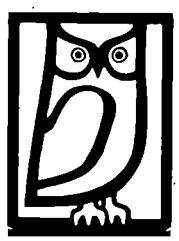
COPYING FOR CLASS USE

The journal gives teachers and their institutions blanket permission for reproducing individual articles for class use in limited numbers (up to 100 copies) at a fee of $1 per copy, payable in advance.

SUBSCRIPTIONS

The journal is published four times a year in January, April, July, and October. The subscription price for 1990 is $115 to institutions, $30 to individuals. Single and back issues will be $30 to institutions and $10 to individuals. The journal is published with the cooperation of the Philosophy Documentation Center. All correspondence regarding subscriptions, renewals, back orders, and related matters regarding the distribution of the journal should be addressed to:

PHILOSOPHY DOCUMENTATION CENTER
Bowling Green State University
Bowling Green, OH 43403-0189 USA



AMERICAN PHILOSOPHICAL QUARTERLY

Edited by
NICHOLAS RESCHER

With the advice and assistance of the Board of Editorial Consultants:

VOLUME 27/NUMBER 4 OCTOBER, 1990

CONTENTS

PUBLISHED BY NORTH AMERICAN PHILOSOPHICAL PUBLICATIONS, INC.
IN COOPERATION WITH THE PHILOSOPHY DOCUMENTATION CENTER
ISSN 0003-0481

AMERICAN PHILOSOPHICAL QUARTERLY

FOUNDED IN 1964

POLICY

The *American Philosophical Quarterly* welcomes articles in English by philosophers of any country on any aspect of philosophy. However, only self-sufficient articles will be published, and not news items, book reviews, critical notices, or "discussion notes" (short or long).

MANUSCRIPTS

Contributions may be as short as 2,000 words or as long as 7,000. All manuscripts should be typewritten with wide margins, and at least double spacing between lines. Footnotes should be used sparingly and should be placed at the end of the paper, numbered consecutively. They should also be typed with wide margins and double spacing. Submissions should always be made *in duplicate*. Only papers whose authors certify that, while under consideration with us, they will not be submitted elsewhere can be considered.

COMMUNICATIONS

Articles for publication and all other editorial communications and enquiries should be addressed to: The Editor, *American Philosophical Quarterly*, Department of Philosophy, University of Pittsburgh, Pittsburgh, PA 15260. Other correspondence not dealing with subscriptions should be addressed to Ms. Dorothy Henle, Operations Manager, *American Philosophical Quarterly*, c/o Department of Philosophy, University of Pittsburgh, Pittsburgh, PA 15260. (For subscriptions, see below.)

REPRINTING

The *American Philosophical Quarterly* regrets that it cannot make reprints available to authors. However, authors have the journal's permission to reproduce limited numbers of their contributions for the use of colleagues and students. (Two copies of the relevant issue will be provided gratis for this purpose.)

While the journal holds the copyright on materials published in its pages, it routinely accords contributors permission to reprint in books or anthologies authored or co-authored by themselves. All other requests to reprint should be addressed to the editor.

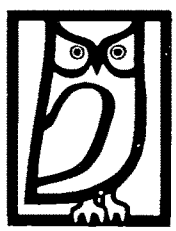
COPYING FOR CLASS USE

The journal gives teachers and their institutions blanket permission for reproducing individual articles for class use in limited numbers (up to 100 copies) at a fee of $1 per copy, payable in advance.

SUBSCRIPTIONS

The journal is published four times a year in January, April, July, and October. The subscription price for 1990 is $115 to institutions, $30 to individuals. Single and back issues will be $30 to institutions and $10 to individuals. The journal is published with the cooperation of the Philosophy Documentation Center. All correspondence regarding subscriptions, renewals, back orders, and related matters regarding the distribution of the journal should be addressed to:

PHILOSOPHY DOCUMENTATION CENTER
Bowling Green State University
Bowling Green, OH 43403-0189 USA



AMERICAN PHILOSOPHICAL QUARTERLY
Volume 27, Number 3, July 1990

BELIEF, CONTRADICTION AND THE LOGIC OF SELF-DECEPTION

Newton C.A. da Costa and Steven French

INTRODUCTION

THE apparently paradoxical nature of self-deception has attracted a great deal of controversy in recent years (for a recent survey, see Mele 1987; a good bibliography is also given in McLaughlin and Rorty 1988). Focussing on those aspects of the phenomenon which involve the holding of contradictory beliefs (where the meaning of "contradictory" in this context will be made clear below), it is our intention to argue that this presents no "paradox" if a non-classical, "paraconsistent," doxastic logic is adopted. Such logics have been extensively developed over the last fifteen years or so (da Costa 1974; Arruda 1980; da Costa and Marconi 1989) and have led to an increased interest in questions of contradiction and inconsistency in general (da Costa 1982; Priest 1987; Smith 1988; French forthcoming). Since the use of such a logic might be regarded as a somewhat radical move in this context, some further remarks are perhaps in order.

First of all, it is our contention that paraconsistent logic gives us a way of modelling a person's internal, inconsistent beliefs, if he or she has any. More than that we would not claim and in particular we are most definitely *not* proposing the overthrow of classical logic in general. Secondly, although we explicitly present, in the appendix, a paraconsistent system which seems capable of accommodating contradictory beliefs, the actual logical details themselves are not essential for the development of our idea (that is, other paraconsistent systems may be equally capable of modelling this phenomenon).

We would also like to emphasise that we are not claiming to have captured all the various aspects of self-deception in terms of such a system, since this would require extensive psychological and philosophical investigation. Rather, we would say that this is a preliminary investigation only, but one that at least captures the apparently contradictory nature of the phenomenon. Put somewhat melodramatically, our aim is to liberate discussion of self-deception from the shackles of a purely classical doxastic logic, thereby permitting a separation of the more philosophical issues from those which might be properly described as "logical."

Finally, we should be clear about what we mean by "contradictory beliefs." There seem to be two interpretations of this phrase which are intertwined in the literature. One is that of holding beliefs in contradictory propositions, which can be represented by $B(p \wedge \neg p)$, where B is the doxastic operator representing "The person x believes that . . . is the case." This, of course, is equivalent to $Bp \wedge B\neg p$. Alternatively, "contradictory beliefs" might be taken to mean the holding and not holding of a belief in a given proposition, or $Bp \wedge \neg Bp$. In what follows we shall reserve this phrase for the former concept, unless we specify otherwise.

As we shall try to argue, these two interpretations are not equivalent in the general case, where inconsistency occurs at the theoretical or "representational" level (in science, say). Consideration of the nature of a theory of truth appropriate for this level leads us to deny the implication from "believing that not-p" to "not believing that p," or $B\neg p \rightarrow \neg Bp$. A belief in a self-contradictory theory, for example, can then be held without implying the holding and not holding of a belief in a given proposition.

At the "factual" level, on the other hand, this implication holds good and the first interpretation above leads to the second. This, we claim, is the level on which the phenomenon of self-deception occurs, to be distinguished from that of self-delusion, which is more appropriately associated with the "representational" level above (these terms will be made clearer in the discussion below).

Thus we shall argue that self-deception means not

only the holding of beliefs in contradictory propositions but also the holding and not holding of a belief in a given proposition. These two conceptions are, however, distinguished in the general case. As we shall see, these considerations are reflected in the structure of the paraconsistent system put forward as a model of the logical aspects of this phenomenon.

The "Paradox" of Self-Deception

Why has self-deception generated so much controversy? In large measure this is because there is no general consensus as to what the phenomenon consists in, making it difficult to state the necessary and sufficient conditions for describing someone as "self-deceived." Numerous attempts have been made to both characterise the state of self-deception and explain how it is generated and maintained (see for example, Mele, *op. cit.*, or the summary in Martin 1986, Ch. 2), all of which have been criticised for failing to capture the full complexity of everyday usage of the term (Martin, *op. cit.*). In what follows, we shall concentrate on those aspects of self-deception which involve doxastic inconsistency or the holding of contradictory beliefs, whilst acknowledging that a full understanding of the phenomenon may require one to go beyond this.

Thus we shall begin with the epistemological "jumping off" point most commonly used in recent discussions: deceiving oneself involves somehow persuading oneself to hold a false belief (not everyone characterises self-deception in this manner; Fingarette, for example, emphasises questions of personal identity (Fingarette 1969) and Martin (*op. cit.*) offers a very general approach in terms of "evading self-acknowledgment of some truth"). The underlying assumption is that self-deception is analogous to "interpersonal deception," where one persuades someone else to believe that p is true whilst knowing, or truly believing, that p is false. According to this model, someone who is deceiving him- or herself must know or believe that p is false, while persuading him- or herself to believe that p is true. Two apparent paradoxes then immediately arise. The first concerns the *state* of self-deception: the above model seems to involve a person simultaneously believing p and believing not-p. Clearly this poses a problem for any view of belief states and their interconnections which assumes, tacitly or otherwise, an underlying classical doxastic logic.

The second paradox concerns the *process* of self-deception: trying to persuade oneself to believe as true what one knows or believes to be false would seem to be an inherently self-defeating task. Mele (*op. cit.*) refers to these as the "static" and "dynamic" paradoxes respectively. We shall be mainly concerned with the former, although we will also have occasion to discuss the latter.

These paradoxes have produced a range of responses (see Mele, *op. cit.*). Those skeptical about self-deception have argued that it is simply impossible. Others (the majority) have claimed that, on the contrary, it *is* possible (usually on the basis of certain paradigmatic examples) and have then gone on to show how the above paradoxes can be subverted or avoided in some way. Thus, some philosophers take the above analogy with interpersonal deception quite seriously and have introduced "divided selves" or some form of mental partitioning. Alternatively, others have argued that this model is inappropriate and have either rejected some aspects of the analogy or abandoned it altogether.

Our aim, in the rest of this section, is to suggest that all of these approaches share a common assumption in an underlying classical doxastic logic. Obviously, we do not have the space here to consider the whole gamut of different responses to the above paradoxes but hopefully, by way of a few representative examples, we can at least indicate the mode of reasoning common to them all. Put bluntly, this takes the form of noting that believing and not-believing some proposition p (or believing p and believing not-p, where these interpretations are distinct in the general case) violates some form of the (classical) principle of contradiction and then proposing some, usually rather convoluted, means of avoiding the paradox by denying that self-deception involves, "in fact," the holding of contradictory beliefs (simultaneously, consciously and in the same mental "division"). The fact that all these approaches can be vigourously criticised in some way or other, then lends support to the claim that perhaps it is the underlying classical "paradigm" that is at fault and that new light might be shed on the problem by placing it in the framework of a logical system which rejects, or weakens, the principle of contradiction. We will further argue that this seems to be particularly appropriate in certain cases where the subject might be said to be in a state of "cognitive conflict."

Let us consider the sceptics' arguments first. These are particularly interesting from our point of view since the sceptic must argue against the possibility of self-deception on the basis of the most general assumptions possible, if he or she is not to be accused of attacking a particular conception only.

Thus Haight, for example (Haight, 1980), begins with the assumption that self-deception is precisely analogous to the case of interpersonal deception, except that it occurs within one person, of course. She then characterises self-deception in terms of someone simultaneously both knowing and not knowing some proposition p, or knowing that p and believing that not-p. However, she argues, this is a contradiction and must therefore be rejected. Furthermore, she notes, interpersonal deception can also be characterised in terms of the deceiver being able to easily bring into consciousness what he or she knows or believes, whereas the deceived is precisely prevented from doing this. Yet, she argues, it would be a contradiction to say that the same person is both able and not able to bring into consciousness the same proposition (thus she covers both the static and dynamic paradoxes).

Haight concedes that it may be possible for a person to be divided, in the sense of there being two, or more, selves in the same body. However, she denies that deception between such selves can be described as self-deception in the literal sense because this requires that one and the same self be both deceiver and deceived. Thus, she concludes, there is no such thing as self-deception in the literal sense.

This line of argument has been criticised for taking the above analogy too seriously. Thus Martin, for example, (*op. cit.*, p. 20), notes that a similar argument can be constructed with the conclusion that there is no such thing as being "self-taught" in the literal sense and concludes that Haight is simply applying the model of interpersonal deception in a misleading way.

From our point of view what is interesting here is Haight's rejection of the possibility of self-deception on the basis of the claim that is leads to a logical contradiction. This comes out even more clearly in a comment she makes elsewhere: ". . . that the idea is self-contradictory and the thing therefore impossible." (Haight 1985, p. 49). There is an obvious, if implicit, assumption here of a classical doxastic logic, in which statements of the form "$Bp \wedge \neg Bp$" are not admitted.

The whole point of the present paper is to suggest that this assumption might be dropped, to be replaced by the adoption of a more appropriate, paraconsistent, logical framework (more appropriate in the sense that such statements can be naturally accommodated). This might then allow one to retain significant features of the above analogy.

Several philosophers have taken this analogy with inter-personal deception very seriously indeed and have attempted to allow for the possibility of self-deception by effectively dividing up the mind. Demos, in his fundamental study of the phenomenon, defines self-deception thus:

> "Self-deception exists, I will say, when a person lies to himself, that is to say, persuades himself to believe what he knows is not so. In short, self-deception entails that B believes both p and not-p at the same time." (Demos 1960, p. 588).

This is possible, he argues, because the self-deceiver is somehow distracted from his belief that p and although aware of it does not attend to it or focus his attention on it.

This, and similar attempts (Pugmire 1969, for example), have been criticised on the grounds that they involve forms of the "dynamic" paradox (Mele, *op. cit.*, pp. 3-4). For someone to be aware of a certain belief but to intentionally ignore it or fail to attend to it, it would seem that that person must already, at some point, have noticed the incompatibility of that belief with others. But then, that person must, at some time, have simultaneously attended to or noticed the contradictory beliefs and, interestingly enough, Demos takes this to be impossible. Again we see the assumption that the simultaneous holding of contradictory beliefs is a logical impossibility.

A more radical tack within this general line of approach is taken by Rorty (1972). Rather than follow Demos, for example, and characterise self-deception in terms of believing both p and not-p, she claims that it obtains if, among other things, a person both believes and does not believe p (p. 394; we shall return to the distinction between believing not-p and not believing that p, below.) However, Mele notes, this

> ". . .straightforwardly violates the principle of contradiction. . . .So if Rorty is not to be charged with a blatant logical error, either "believes" is being used equivo-

cally or the agent must somehow be split." (Mele, *op. cit.*, p. 4).

Rorty's choice is to argue for a pluralistic conception of the self, according to which one person in fact constitutes a multitude (*op. cit.*, p. 404). This is clearly too much to swallow for many people and again the criticism has been levelled that Rorty's conclusion is simply a case of taking the analogy with inter-personal deception too far (see, for example, Bok 1980).

Leaving aside such counter-arguments, we see yet again the role played by the classical principle of contradiction in reaching the above conclusion. The assumption of a classical doxastic logic is brought out quite sharply by Mele's charge that violating the principle of contradiction would be a "blatant logical error." In this context it is interesting to note Szabados' assertion that rather than demystifying self-deception, as she intended, Rorty simply reinstates the (static) paradox by characterising the phenomenon in the way she does (Szabados 1974). Unless self-deception is to be regarded as merely a form of split personality on the above approach, the "multitude" of selves must presumably be integrated in some way so that they can be said to constitute a "whole" person. But then, given such integration, it is not clear why the supposed "paradox" is not simply reinstated; that is, it is not clear why we cannot attribute contradictory beliefs to that person, considered as a person. Of course, the mental "divisions" could always be drawn exactly so as to separate the contradictory beliefs, but this seems somewhat ad hoc.

A further assumption that is made in the above accounts is that self-deception is intentional. Mele has recently defended the idea that self-deceivers typically do not intentionally deceive themselves, thus breaking away from an important aspect of the inter-personal model (Mele 1983; *op. cit.*, pp. 9-10). According to this approach, self-deception essentially arises through the operation of a person's desires for certain beliefs. Such desires then lead to a manipulation of the data relevant to the truth-value of the objects of those beliefs. Thus,

> ". . .because, e.g., he (the subject) takes a certain datum *d* to count against *p*, which proposition he wants to be the case, he may intentionally or unintentionally shift his attention away from d whenever he has thoughts of *d*; but to do this he need not believe that *p* is false." (1983, p. 372).

The requirement that a self-deceived person believes not-*p* (regarded as the truth) is therefore dropped on this view and there is no holding of contradictory beliefs.

The core of this position lies in denying that a self-deceived person knows the truth, this knowledge being blocked off by some kind of manipulation of the data relevant to the truth value of *p*. However, this kind of picture, although plausible in some cases, is incapable of accommodating certain others, which might fairly be described as "paradigmatic" examples of self-deception. Thus, let us consider the well-known example of a close friend or member of the family suddenly dying and the bereaved asserting, "I know that x is dead but I just don't believe it" (see, for example, Glick, Weiss and Parkes 1974, p. 54; or the discussion in Martin, *op. cit.*, pp. 23-24). Such a person might not only assert this belief but might further reveal it through his or her behaviour, such as continuing to lay a place for the dead person at the dinner table, for example. In such cases, it would seem that we are entitled to say that the relevant knowledge is typically present in the person's awareness, and is therefore conscious, the conflict with his or her belief leading to severe cognitive discomfort, which in turn is manifested in terms of extreme or bizarre forms of behaviour.

Further examples can be drawn from cases of what are described as "transient situational personality disorders" (Ullman and Krasner 1969, pp. 315-28) where a person exhibits certain acute reactions, usually temporary or transient, in response to what is perceived as an overwhelming situation. "Overwhelming" situations are those which are difficult to cope with because they are completely at variance with prior experience (*ibid.*, pp. 316-19). With nothing to go on, a person in such a situation may act in an erratic or uncoordinated way or ". . .may emit parts of two inconsistent operant behaviors at the same time." (*ibid.*, p. 319). The end result is confusion, unreasonable behaviour and the loss of ability to make discriminations which the person was previously able to make.

It is this which suggests that such a person is in a genuine state of conflict which is expressed through inconsistency between forms of behaviour or between such behaviour and verbal assertions. The

person is unable to choose between two conflicting and contradictory sets of beliefs and it is this inability which characterisations such as Mele's above seem unable to capture.

Mele himself argues against this kind of interpretation on the grounds that if a person consistently acts as one would act who does not believe that *p* then we have excellent reason to deny that such a person also believes or consciously knows p (*op. cit.*, p. 373). That is, he seems to be saying that, in the case above, if *x* asserts that he or she knows that *y* is dead but continues to act otherwise, then we should accept that *x* has somehow pushed the knowledge out of conscious awareness, or effectively blocked it off by denying the implications of the relevant data, and really does believe only that *y* is not dead. There is, therefore, no conflict.

However, this strikes one as rather implausible. How can *x* somehow shift his or her attention away from this knowledge and yet still assert that he or she knows *y* to be dead? The assertion seems to exactly express conscious awareness of the relevant knowledge. Furthermore, our brief discussion above suggests that the person's actions will not, in fact, be consistent and Mele gives no justification for choosing the behaviouristic construction of "*x* believes that *p*" over the disposition-to-explicit-assent or commitment-to-assent ones. As Martin notes, referring to people in such situations,

> ". . .the ordinary concept of belief seems sufficiently flexible to allow us to say they irrationally believe what they know to be false, and this belief explains the conflicting emotions and behavior. And while pathological forms of self-deception are involved in these . . . examples, there is no reason to deny that less disturbed self-deceivers could know the truth while believing the opposite." (*op. cit.*, p. 24).

Finally, claiming that there is, in fact, no conflict here clearly runs counter to not only our own intuitions but also those of the psychologists, who would be inclined to treat such a person as someone requiring some degree of psychiatric counselling, especially if the condition were to persist. In other words, anyone making such a claim might well be accused of not taking such cases seriously, psychologically speaking.

Nevertheless, Mele's type of approach does seem to work well in other kinds of cases. Let us take the commonly used example of a person *x* who is presented with evidence that his or her partner *y* is being unfaithful but refuses to accept such a possibility and continues to believe in *y*'s fidelity. Such a belief may and generally will, be manifested both verbally and behaviourally, the absence of any assertion of conscious knowledge of the contrary facts reflecting the absence of mental conflict. In such cases, the idea of *x* shifting attention away from the relevant datum seems quite plausible.

What, then, distinguishes such cases from that of the bereaved person above? One significant difference, as we have tried to indicate, is that in the latter case but not the former there is some "cognitive discomfort" arising from the confrontation between contradictory beliefs. Thus we might label as cases of "self-deception" those cases where some such cognitive discomfort is manifested, taking it to reveal an inner doxastic conflict, and reserve the term "self-delusion" to describe the others.

This suggestion is broadly in tune with Graham's characterisation: "*S* is self-deceived just when *S* believes what *S* desires to believe and does so when confronted with contrary facts." (1986, p. 225), where this confrontation involves *S* perceiving or knowing these contrary facts. Thus, ". . .self-deceivers see contrary facts and appreciate that they are contrary to their belief. They see them as contrary." (*ibid.*, p. 226).

It is this aspect of conscious knowledge which is captured by Moore's paradox, which any doxastic logic must face: "*S* knows that p but does not believe it" (da Costa and French 1989a). And it is this, in turn, which is exemplified by some of the extreme, or "true," cases of self-deception as noted above.

However, this "cognitive conflict," springing from conflict states involving contradictory beliefs, is not manifested when someone is self-deluded. As Graham puts it:

> "Someone who is self-deluded believes what he wants when confronted by contrary facts and in some unconscious sense appreciates that the facts are contrary, but this appreciation never occurs in consciousness." (*op cit.*, p. 277).

In other words, delusion mechanisms of various degrees of strength operate to hide the contrary data from the person's conscious awareness, with the consequence that no "cognitive discomfort" results.

Of course, one may experience varying degrees of cognitive discomfort, depending on the situation,

and as the strength of the underlying conflict diminishes, self-deception may shade over into self-delusion (cf. Graham, *op. cit.*, p. 229; Mele, *op. cit.*, p. 14). The extent to which a "contrary fact" or falsifying piece of evidence can be shifted out of one's immediate conscious awareness may depend not only on the strength of one's desire to cling to an erroneous belief, but also on the "epistemic weight" or "force" of this fact or piece of evidence. One can easily imagine situations in which the "weight of evidence" gradually builds up, until it is too much to bear and the conflict is forced into the open. Delusion mechanisms, such as shifting attention, misinterpretation and one-sided evidence gathering (Mele 1983) may simply collapse under the epistemic weight of "the facts," allowing the two contradictory beliefs to collide into one another. Of course, many people will then accept the evidence and drop their false belief. However, the desire to retain the latter may be so strong and so firmly supported by other reasons, including, perhaps, psychological ones, that both beliefs continue to be held and the person is said to be in a state of self-deception.

These "other reasons" may, of course, include evidence gathered at an earlier time, regarding the previous behaviour of a loved one, for example. However, one must exercise obvious caution in allowing for conflicting pieces of evidence in this context. The situation in which one has (different) evidence for both p and not-p is commonplace in science and to say that scientists are generally deceiving themselves, in the literal sense, is obviously stretching our understanding of the term too far, to put it mildly! A significant difference between this situation and that involving self-deception is as regards the epistemic attitude of the persons concerned. Whereas someone who is self-deceived can be said to have a "factual" belief in a proposition which is regarded as true in the strict or correspondence sense, scientists are better characterised as holding "representational" beliefs in their theories, regarded as "approximately" or "quasi-" true only (French forthcoming). We shall return to this distinction below. (Of course, a scientist may still become self-deluded and an investigation of historical examples of this phenomenon might shed some light on the kinds of delusion mechanisms involved).

Returning to the distinction between self-deception and self-delusion, we can see that, in arguing that "true" cases of the former involve states of conflict centering around contradictory beliefs, we are also arguing, against Mele (1987, p. 10), that a self-deceived person must know "strong evidence" against the proposition he or she falsely believes. The evidence must be precisely strong enough to overcome the mechanisms of delusion and force the contradictory beliefs into conscious awareness. This can obviously occur not only through the gradual accumulation of data but also through the discovery of a particularly significant or "weighty" piece of evidence which dramatically "tips the scales," as it were. Of course, it can be argued that this is too strong, that someone can become self-deceived through the mechanism of ignoring significant data (Mele 1983, p. 375; 1987, p. 9) but this we would claim, is a case of self-delusion, rather than self-deception *per se*.

Our aim in this section was to point out, through a selection of representative examples, the assumption that is usually made in discussions of this problem, namely that of an underlying classical logic of belief. It is this which impels people to reject the possibility of the literal holding of contradictory beliefs and leads them to construct various elaborate schemas for evading the apparent "paradox," some of which verge on the baroque. Given that these attempts differ so dramatically, that, relatedly, there is no consensus as to how self-deception should even be characterised and that all the numerous proposals which have been put forward are open to criticism of one form or another, the suggestion that perhaps it is the underlying logical framework which needs to be changed gains a certain plausibility. Of course, this is not to exclude the possibility that consensus might be reached, a general approach might be agreed upon and some account of self-deception constructed which is both resistant to the criticisms levelled against previous attempts and capable of accommodating all the paradigmatic examples of the phenomenon. We merely wish to suggest that, given the confused state of the field, it might be worthwhile to consider abandoning the underlying classical framework and looking at some of the alternatives.

Furthermore, as we have indicated, there exist clear examples of self-deception which seem to cry out for treatment in terms of the holding of contradictory beliefs. At the very least, this would seem to be a more natural way of treating such cases than the introduction of notions of attention shifting, ignor-

ing evidence etc. The latter, we have argued, are more appropriately employed in understanding cases of self-delusion, where cognitive conflict does not arise. The only impediment to the former approach is the classical law of contradiction and dropping this, by way of some paraconsistent system, liberates us from the restrictions imposed by a particular form of logic and opens up previously unexplored possibilities.

Elsewhere, we have outlined some of these possibilities in terms of logical systems capable of accommodating certain forms of contradictory beliefs (da Costa and French 1989a) and a related system is given in the appendix to the present paper. Before we discuss paraconsistent logic, however, we need to examine some of the different ways in which inconsistency arises within a system of beliefs and to consider how these bear upon the problem of self-deception.

The "Hobgoblin" of Consistency

The title of this section is taken from Emerson: "a Foolish consistency is the hobgoblin of little minds, adored by little statesmen and philosophers and divines.", quoted in Kyburg (1987). A number of philosophers, in recent years, have come to question this "adoration" of consistency and have argued against the stringent imposition of the requirement that inconsistency be avoided at all cost (in addition to the authors cited in the Introduction, see also Kyburg, *ibid.*; Harman 1986; and Cherniak 1986). This work can be located within a more "naturalistic" approach in general, which is capable of accommodating certain forms of contradictory or inconsistent beliefs. However, a principal concern of these authors is the relationship between consistency and rationality or, more generally, between logic and reasoning and, although we will touch on this, we are more interested in the kinds of inconsistency which can arise and whether these involve self-deception or not.

Let us begin with the example of someone who believes both p and q, where p implies not-q. In this case, the "contradiction" may be said to be indirect (where the term "contradiction" is used in a loose sense), leaving open the possibility of some kind of reconciliation of the jarring beliefs, depending on the nature of the implication concerned (cf. Martin, *op. cit.*, p. 23). Suppose, however, that no such reconciliation is possible, what can we say about this kind of inconsistency?

First of all, we must clearly distinguish between a person being aware and unaware of the implication and, if the latter, between being intentionally or purposefully unaware and being unintentionally so. A person may obviously be unintentionally, that is completely and absolutely, unaware of an implication which ultimately results in a contradiction with other held beliefs. We are not logically omniscient, in the sense of being able to immediately deduce all the consequences of a given proposition that *can* be deduced and anyone who demands such omniscience is clearly asking too much. There are many trivial implications of a given set of beliefs which would simply clutter up one's mind if added to that set (Harman, *op. cit.*, pp. 12-15). Moreover, our minds are capable of accommodating only a finite set of explicit beliefs and therefore some limit on the number of implications we can draw is forced upon us (explicit beliefs are not necessarily conscious beliefs: a belief may be explicitly represented in one's mind and figure in an explanation of one's behaviour, yet be unavailable to one's consciousness; Harman, *ibid.*, p. 14).

Whether an implied belief is deemed trivial or not depends on the context, of course, but even non-trivial implications might escape us since we are not even "locally" omniscient. It may take a complex and lengthy proof for one to become aware of a certain implication of one's beliefs; that is, the implication may not be, in any sense, obvious. This can be true of both explicit and implicit beliefs. Such a situation can be expected to occur quite frequently and examples can be given from the history of science, for example, where certain logical consequences of a given hypothesis were not immediately perceived when it was first proposed and their later discovery contributed to the hypothesis's degree of confirmation (Garber 1983; da Costa and French 1988).

The degree to which a given implication can be said to be "obvious" clearly depends on both the length of the deductive chain and the number of beliefs involved. Someone who is apparently unable to see, or appears to be unaware of, a particularly simple or "obvious" implication of their set of beliefs might be said to be logically incompetent rather than to be guilty of some more serious doxastic misdemeanour (the person's lack of awareness

being revealed through the usual means of behaviour, disposition to assent, etc.).

Once we become aware of a certain implication and follow the consequences of a given belief or set of beliefs, we might discover, of course, that our chain of reasoning leads us to a belief which is inconsistent with one or more others that we hold (that is, the belief directly contradicts these others). Given what has been said, that we are not logically omniscient, there may be many "implied inconsistencies within our overall corpus of belief. The demand that we conduct a straightforward check for such inconsistencies and suitably adjust our system of beliefs once they have been uncovered, is clearly hopelessly unrealistic, given both our lack of omniscience and the length of time involved: Cherniak has estimated, on very reasonable assumptions, that a consistency test for a modest system of only 138 beliefs would take more time than the currently accepted age of the universe! (Cherniak *op. cit.*, p. 756).

It should not, therefore, come as too much of a surprise to discover the presence of "implied inconsistencies" within one's system of beliefs. Indeed, the existence of inconsistency within science and even logic and mathematics is now gaining widespread recognition (Norton 1987; Smith, *op. cit.*; Cherniak, *op. cit.*). The crucial questions, however, are how one regards the inconsistent beliefs and what one does with them.

One possibility is to become intentionally or purposefully unaware of the implied contradiction by, for example, shifting attention away from the belief or beliefs at the head of the deductive chain, or shifting attention away from the chain itself, or employing some other delusion mechanism. The longer or more complex the chain of reasoning involved, the easier it is to shift attention away from it, put it out of focus, or whatever. Someone who does this can fairly be described as self-deluded.

Suppose, however, that one does not block off the implication in this manner but accepts it fully and completely into one's awareness, what then? Assuming that the implication is valid and that the implied or indirect inconsistency therefore carries as much epistemic force as a direct one, it might be argued that the only rational thing to do is to suitably rearrange one's set of beliefs with a view to eliminating the contradiction from the system. A failure to do so would then bring down an accusation of self-deception.

However, it may be extremely difficult, if not impossible, to eliminate the contradiction by simply removing one of the contradictory beliefs from the belief system, because this belief may, and in general almost certainly will, be connected in many different ways with a whole range of other beliefs, themselves ineliminable perhaps, practically speaking. The removal of inconsistency in this manner may therefore be a practically impossible, or near impossible, undertaking. One might have good reason to hold both of a pair of contradictory beliefs. Thus, for example, the same empirical evidence might equally support two conflicting theories, or different, but equally acceptable, pieces of evidence might support two contradictory propositions within a given theory, such as Bohr's theory of the atom, to name one example. As we have already said, it would clearly be going too far to accuse Bohr, or any other scientist in a similar situation, of self-deception. However, as we shall see, a simple consideration of scientists' epistemic attitudes with regard to their theories, inconsistent or otherwise, will shed some light on the difference between this kind of situation and that of someone who is in a state of self-deception.

One way of dealing with an inconsistency, at least temporarily, would be to divide one's overall set of beliefs into consistent sub-sets (Smith, *op. cit.*; Kyburg, *op. cit.*). By deriving the implications of these sub-sets one can make many inferences from the inconsistent set, without being committed to the deductive closure of the latter (this proposal is therefore related to our considerations about logical omniscience above). Indeed, abandoning the imposition of deductive closure might be forced upon us if the problem we are dealing with is found to be "computationally intractable" in the sense that there is an exponential explosion of operations involved (Cherniak, *op. cit.*). Consideration of the way in which evidential support bears upon the implications drawn from the consistent subsets may then indicate how a consistent alternative to the original inconsistent set of beliefs might be constructed. Smith, for example, has recently illustrated the heuristic role such consistent sub-sets play with regard to the occurrence of inconsistent theories in the history of science (Smith, *op. cit.*, 1988b).

There is an interesting analogy here with the "di-

vided selves" proposal for characterising self-deception. Thus a self-deceived person might be regarded as someone who holds two mutually inconsistent but internally consistent sub-sets of belief and is capable of drawing inferences from each. However, the analogy cuts two ways. On the one hand, it might be used to press the case against this characterisation on the basis of the argument that scientists regularly perform this type of division and could not, or should not, be regarded as self-deceived. On the other, as with the divided selves proposal, it might be questioned whether this division into consistent sub-sets can be made in any clear and non-ad hoc manner. Surely there must be some degree of integration across the various sub-sets for them to be considered *as* sub-sets of the original theory. Isolating the units, or sub-units, of a theory that are necessary to deduce a given evidential statement is a well-known problem in the philosophy of science and it is interesting to note that Waters has offered a possible solution to this problem by introducing relevance logic into the hypothetico-deductive approach (Waters 1987).

Since relevance logic can be considered a form of paraconsistent logic (see below) this brings us on to a second way of dealing with inconsistent beliefs, that of introducing an underlying paraconsistent logic of inference (whether deductive or inductive). As we have already said, several times, and as we shall discuss below, such logics are capable of accommodating inconsistency by weakening or abandoning altogether the classical law of contradiction. Although Smith and Norton (*op. cit.*) have argued that the logic of scientific inference is not paraconsistent, tacitly or otherwise, this is merely a contingent matter at best and given that the central thesis of this paper is that states of self-deception might be easily and naturally modelled through an underlying paraconsistent doxastic logic, something must be said about what distinguishes such states from the epistemic states of scientists with regard to inconsistent theories, which might be modelled in the same way.

To get the discussion rolling, let us consider Priest's recent defence of "dialetheism": the view that there are *true* contradictions (the emphasis on truth here is important, as we shall see). He begins (Priest 1987) by considering the well known semantic and set-theoretic paradoxes and argues, forcefully and convincingly, that these provide prima facie examples of such true contradictions. He then goes on to claim that it is possible to hold inconsistent or contradictory beliefs, essentially on the basis that our grounds for attributing belief to someone are sufficiently broad and flexible as to allow us to attribute contradictory beliefs to them; there are various actions, for example, which can be quite naturally connected with inconsistency in a belief set (*ibid.*, Ch. 7). Finally, he argues that it may be *rational* to hold contradictory beliefs, drawing on examples from the history of science to support his conclusion that, ". . . whatever kind of argument it takes to make something rationally acceptable, an inconsistency can have it." (*ibid.*, p. 127).

However, at this point Priest's reasoning seems to have gone awry. It is highly implausible, if not absurd, to maintain that Bohr regarded his inconsistent theory of the atom as *true*. Indeed, given the frequency of theory change in the history of science, it would seem to be implausible to regard *any* scientific theory as true, in the literal sense. Rather they should be regarded as, at best, "approximately" or "quasi-" true only, (the empiricists would have us drop truth entirely at the theoretical level, to be replaced with something like "empirical adequacy," for example).

One of the problems with the notion of "approximate truth" is its perceived lack of formalisation. However, a solution to this problem has in fact been put forward via the introduction of "partial structures" within the model-theoretic approach in general (Mikenberg, da Costa and Chuaqui 1986). It is not our intention to enter into any of the technical details of this idea (summaries can be found in da Costa and French forthcoming a and 1989b) suffice to say that a "partial structure" can be thought of as a partial model of a given domain which captures only some of the relations between the elements of that domain and thus reflects the incomplete state of our knowledge at any given time. This can then accommodate belief in, and acceptance of, a scientific theory as "pragmatically," "approximately" or "quasi-" true only. Within the partial structures one can identify "empirical sub-structures" which model the observable aspects of the domain concerned and which can be regarded as true in the literal or correspondence sense. (The question immediately arises, of course, whether the empirical sub-structures can be clearly and distinctly marked off from the theoretical structures in which they are embedded. Although we believe that they can, in at least a number

of important cases, it is obviously not easy to lay down the conditions for such a demarcation; see van Fraassen 1985).

There is an interesting analogy here with Sperber's distinction between "representational" and "factual" beliefs (Sperber 1982). Briefly put, the latter relate to propositions accepted as true in the correspondence sense, whereas the former are "opinions," "convictions," or beliefs in general. The difference is that in the case of a factual belief there is awareness only of a fact, whereas in the case of a representational belief there is a commitment to a certain representation. In particular, this representation might be "semi-propositional" in the sense that it fails to identify one and only one proposition, unlike propositional representations which do exactly this. A semi-propositional representation can be given as many propositional interpretations as there are ways of specifying the conceptual content of its elements (for more on the analogy between partial structures and empirical sub-structures and semi-propositional and propositional representations, respectively, see French forthcoming; da Costa and French forthcoming b).

Again, we shall not go into all the details here. What is crucial for the present discussion is that the epistemic attitude of scientists towards their theories is weaker than that of belief in, or acceptance of, them as *true*, literally, universally or in the correspondence sense. This can be captured in terms of partial structures, or models, with the theories regarded as "approximately" or "quasi-" true only. The weaker epistemic attitude can then be characterised as holding a representational belief in the theory concerned; that is, a belief in the theory as partially true only. "Scientific" beliefs are *not* factual beliefs in the sense of a belief in a theory as a literally true representation of the way the world is, at least not at the "theoretical" level (beliefs concerning the empirical sub-structures may, however, be regarded as "factual" beliefs).

We realise that we have introduced a number of unfamiliar terms here and that there is a lot more that needs to be said on these matters. Further details are provided in our other papers, already cited. All that is really important for our present purposes is to acknowledge that the epistemic attitude of a scientist with regard to a particular theory is weaker than that of belief in that theory as true in the correspondence sense. Once we allow this, the door is opened to a coherent treatment of inconsistency in science which can account for the rationality of believing in a contradictory theory, for example, as partially or "pragmatically" true only (French, *op. cit.*). Thus, a scientist might "entertain," to use Smith's vocabulary (Smith 1988a), or "provisionally" or "pragmatically" accept, to use ours, two or more mutually inconsistent theories, each equally supported by the available empirical evidence, with the tacit understanding that further research will indicate which is closer to "the truth" (for more on the underdetermination of theories, see Pereira and French, forthcoming). Equally, an "internally" inconsistent theory, such as Bohr's theory of the atom, might be provisionally accepted, as heuristically fruitful at least, if the internal contradiction is outweighed by other factors, such as, and principally, evidential support for the theory reflecting the extent to which it does in fact model its intended domain (we suspect that this was precisely the underlying epistemic situation in Bohr's case, for example). In either case, it is rational to accept such theories, but only in a provisional sense, that is, as partially, approximately, pragmatically, or whatever, true only.

However, the same can obviously not be said for "factual" beliefs or beliefs concerning the "empirical sub-structures." In *this* case, the supporting evidence for the belief that p can be regarded as evidence against the belief that not-p, where "belief that p" is understood as shorthand for "belief that p is true in the correspondence sense." It is then irrational to hold inconsistent beliefs because, following the correspondences, there will be implied inconsistencies in the "facts" or at the level of observable phenomena. Acting on such inconsistent or contradictory factual beliefs will then lead to conflicting or contradictory behaviour, suggesting, to an observer, that the person concerned is in a state of genuine doxastic conflict. And it is precisely in this kind of situation that a charge of self-deception, as characterised above, seems most appropriate (cf. Williams, *op. cit.*).

The upshot of our discussion, then, is that what distinguishes the self-deceiver from the epistemically cautious scientist, say, is that the former holds inconsistent "factual" beliefs in propositions regarded as true in the literal, correspondence sense. Moving away from this conception of truth then allows a rational accommodation of inconsistent beliefs at the theoretical or "representational" level. At

this level, unlike that of the "factual" or observable, belief goes beyond the evidence, of course. Denying the weight of this evidence, for some reason, might induce a state of self-delusion but this is not the same thing as self-deception, as we have repeatedly emphasised.

This focus on truth also helps to illuminate the distinction, noted above, between believing that not-p and not believing that p. Rorty appears to regard these as distinct states (Rorty, *op. cit.*) as does Martin (*op. cit.*, p.21), although there is little or no discussion as to why they should be distinguished. The perception that they may amount to the same thing can perhaps be traced back to the idea that belief is exclusive, something which Mele, for example (Mele 1987, p. 4), regards as a special feature of believing which distinguishes it from, say, desiring. But what is it about belief, as such, that makes it exclusive? Is this just some inherent unanalysable feature of the state or can it be traced back to something else? Elsewhere (French, *op. cit.*) it has been argued that there is a strong connection between the exclusive nature of belief and the often made linkage between belief and (correspondence) truth.

Thus, in formal terms, a belief in a contradiction can be written:

$$B(p \wedge \neg p)$$

or, under natural presuppositions,

$$Bp \wedge B\neg p$$

where B is the doxastic operator again. If believing that not-p implies not believing that p, that is, $B\neg p$ implies $\neg Bp$, then the above will give:

$$Bp \wedge \neg Bp$$

Hence the inconsistency "spreads" from the propositions to the belief states themselves and holding a contradictory belief is equivalent to holding inconsistent beliefs.

Now consider truth. On the standard account, one usually accepts that:

$$T\neg\alpha \rightarrow \neg T\alpha$$

In other words, truth is also regarded as exclusive. If we further accept that "belief that p" is shorthand for "belief that p is true," then we obtain:

$$BT\neg\alpha \rightarrow B\neg T\alpha$$

which informally seems to imply that:

$$B\neg T\alpha \rightarrow \neg BT\alpha$$

Dropping the T's, we then have:

$$B\neg\alpha \rightarrow \neg B\alpha$$

Thus the exclusive nature of belief seems to follow from that of truth, together with the commonly made connection between the two. We might say that there is nothing inherent in belief itself which causes it to be exclusive; it's all in the propositions and whether they are regarded as literally true or false or not.

Dropping this connection between belief and truth and introducing some weaker epistemic attitude, as discussed above, then allows us to deny the above equivalence and block the "spread" of inconsistency. A scientist working with an inconsistent theory can therefore be said to believe in a (self-) contradictory theory (as partially true) but not to hold inconsistent beliefs. A logic of science suitable for accommodating such a situation should thus include, as a principle:

$$\neg(B\alpha \wedge \neg B\alpha)$$

In (da Costa and French 1989a) a form of paraconsistent doxastic logic is presented which can accommodate:

$$B(\alpha \wedge \neg \alpha)$$

but which has the above principle as a theorem. As we noted there, this means that when we reason about our own beliefs, our "external" logic is classical, whereas our "internal" logic is paraconsistent (cf. Rescher and Brandom 1980, §26, who claim that even if our "object" theory is inconsistent, our "meta-theory" should be consistent; also cf. Priest, *op. cit.*, who denies the distinction between "object" and "meta-" theory but accepts the exclusion principle for truth on the basis of the rather vague claim that contradictions should not be multiplied beyond necessity; for criticism, see da Costa and French forthcoming c).

Abandoning the above principle then allows us to accommodate statements having the form of "Moore's Paradox" (da Costa and French 1989a) and inconsistent factual beliefs in general (or factual beliefs in contradictions, since without this principle it seems natural to assume that $B\neg\alpha$ implies $\neg B\alpha$). Since these are precisely what characterise the state of self-deception in our view, the conclusion we reach is that believing that not-p and not believing that p should not be regarded as distinct where self-

deceivers are concerned and, further, that a doxastic logic suitable for formalising such belief systems should not contain the above theorem.

In other words, belief itself is not necessarily exclusive. It only appears so when the usual connection is made between belief and truth as correspondence, with exclusivity "flowing" from the latter. Loosening this connection by acknowledging the possibility of epistemic attitudes weaker than "belief that p is true," then allows us to distinguish believing in a contradiction, or contradictory theory, from holding inconsistent beliefs. These "weaker" attitudes, whether they be "believing a theory to be partially true," or "pragmatically accepting" or "entertaining" a theory, fall under the rubric of, or are related to, what we have called, following Sperber, "representational beliefs," where the content of the corresponding proposition, or possible propositions, "goes beyond" the evidence.

However, the above is not the case for "factual" beliefs, where the representation concerned precisely represents, isomorphically perhaps, the evidence or "facts." Here the appropriate attitude is that of "belief in p as true in the correspondence sense" and this *is* exclusive. The above distinction therefore cannot be maintained in the case of factual beliefs, which are exactly those involved in examples of what we have called self-deception. States which appear similar to those of self-deception via the manifestation of similar "symptoms," but for which it is possible to distinguish belief in not-p from non-belief in p (in the sense that belief in not-p does not imply non-belief in p) should, we suggest, be more appropriately characterised as examples of self-delusion. Shifts of attention away from, or manipulation of, relevant evidence are clearly more likely to occur when there is some epistemic room to manoeuvre, in the sense of a certain looseness of fit between the data and the representation, as in the case of representational beliefs. There is simply no epistemic purchase for such delusion mechanisms where factual beliefs are concerned, since the relationship between the data and the representation of this data, or the empirical sub-structure, can be described as isomorphic. (This can obviously be disputed and, indeed, the modelling of empirical evidence is rather more problematic than is commonly supposed. However, even if a case could be made for claiming that delusion mechanisms can exist at the level of factual beliefs also, we would still argue that not all cases of apparent doxastic conflict at this level can be interpreted as examples of self-delusion.)

It might be remarked that the examples of self-deception which we have mentioned should be analysed in terms of knowing and not-believing p, rather than believing and not-believing (Martin, *op. cit.*, pp. 23-24). However, given that on all conceptions of what knowledge is, to know p is to hold a belief that p (which belief may then be taken to be warranted, true or whatever), a doxastic contradiction still arises. Furthermore, as is well known, various logics of knowledge can be constructed (see Hintikka 1962; Rescher 1968) and these can also be subjected to a paraconsistent treatment.

The above can be regarded as a preliminary attempt to locate the problem of self-deception within the context of inconsistency in general. Let us now return to the central thesis of this paper and the question as to how such inconsistencies might be formalised.

Paraconsistent Logic[1]

Let T be a deductive theory whose language has a symbol for negation. T is said to be inconsistent if the set of its theorems contains at least two formulas or sentences, one of which is the negation of the other; otherwise T is consistent. T is called trivial if the set of its formulas (or sentences) coincides with the set of its theorems; otherwise T is called non-trivial. Most common logical systems, such as the classical and intuitionistic, for example, do not separate inconsistency from triviality; that is, a theory based on such logics is trivial if and only if it is inconsistent. Paraconsistent logics can serve as the underlying logic of theories that are inconsistent but non-trivial. Thus, in general, a logic can be called paraconsistent if it contains, or can be the logic of theories which contain, sets of theorems which are incompatible with classical logic, in the sense that if they are added to classical logic a contradiction results.

In a broad sense, a paraconsistent system can be said to result if the scope of the principle of contradiction is restricted in some way. By the principle of contradiction one normally means one of the following (which are not, however, equivalent):

i) of two contradictory propositions, one of which is the negation of the other, one is false;

ii) $\neg(\alpha \wedge \neg \alpha)$, where $\neg$ and $\wedge$ stand for negation and conjunction respectively;

iii) a predicate cannot simultaneously belong and not belong to the same subject (cf. Orwell's "doublethink," presented by Martin, *op. cit.*, as a possible form of self-deception).

The history of such logics can be traced back to Lukasiewicz and Vasil 'ev who, between 1910 and 1911, independently argued for the revision of Aristotelian logic via the elimination of some form of the principle of contradiction (for a history of paraconsistent logic see Arruda, *op. cit.*, and 1989). Although Jaskowski constructed the first paraconsistent propositional calculus in 1948, the development of this logic in its present form can be traced to the work of da Costa who, from 1954 onwards, independently constructed several such systems at both the propositional and predicate levels, as well as the corresponding calculi of descriptions and several applications in set theory. The last fifteen years or so have seen an explosive growth of interest in the subject, as the principles and internal structure of the various systems have been refined and developed, and many wide-ranging applications discovered (for recent discussions, see Priest and Routley 1984; da Costa and Marconi 1989; Priest, Routley and Norman 1989).

As the field stands today, however, three broad types of paraconsistent logic can be distinguished (Smith 1988a, p. 245; for an alternative classification in terms of "non-Scotian" systems, see Marconi 1981):

1) those based on da Costa's technique of supplying a non-standard semantics for negation (da Costa 1974);

2) so-called "non-adjunctive" logics (Rescher and Brandom 1980);

3) relevant logics (Routley, *et. al.*, 1982).

Taking the last first, the principal motivation for relevant logic is to overcome the well known "paradox of implication:"

$A \rightarrow (B \rightarrow A)$

This is achieved by introducing a restriction on the consequence relation such that it holds only if there is a propositional variable shared between the conclusion and a premiss. Thus inferences from a contradiction to any proposition whatsoever are disallowed on this approach. Although the original motivation of relevant logic had nothing to do with paraconsistency, the definition of an implication relation weaker than classical material implication can only be justified, in pragmatic terms, by reasons tied to some paraconsistent position. Against this approach it has been objected that the formal treatment of the logical constants is in no way related to the body of linguistic practices within which the constants are supposed to receive their intuitive meanings (Moriconi 1981). This seems a particularly telling criticism against a position whose main "raison d'etre" is as a response to the counter-intuitive nature of classical implication.

As their name suggests, non-adjunctive logics reject the rule of adjunction:

$$\frac{A, B}{A \wedge B}$$

Consequently, a contradiction such as $A \wedge \neg A$ cannot be true, although both its "sides," A and $\neg A$, can be true. (A non-adjunctive doxastic logic would therefore, presumably, reject the principle that to believe or accept each of a number of beliefs is to believe/accept their conjunction; cf. Kyburg 1961). However, these systems have been criticised for not having a valid multi-premiss inference which is regarded as implausible for systems which are intended to model practical inference (Priest and Routley, *op. cit.*, p. 7).

One response to this problem is to allow a certain amount of conjoining of premisses, up to "maximal" consistency (Schotch and Jennings 1980). In this way the set of formulas of the theory is effectively divided up into a number of internally consistent "partitions." This is clearly analogous to the method of dealing with inconsistent beliefs by splitting them up into consistent subsets. As Smith remarks, this approach

> ". . .embrace(s) within a single logistic system the process of hypothetical reasoning from the *consistent subsets* of an inconsistent set of statements. For example, in Rescher and Brandom's non-adjunctive logic, the deductive closure of an inconsistent set of statements is simply the union of the classical deductive closures of its consistent subsets. Identifying this set is simply a matter of hypothetical reasoning under classical rules of deductive inference." (*op. cit.*, p. 245).

In such a division the coherence and integral nature of the overall theory or set of statements/beliefs is lost (cf. our comments with regard to the analogous

approach towards self-deception above; also see Priest and Routley, *op. cit.*, p. 8).

Finally, we come to da Costa's "positive plus" approach, so called because such systems retain the whole of classical positive logic, but allow negation to behave non-classically. The most well known paraconsistent logics of this type are those based on the propositional calculi C_n, $1 \leq n \leq w$, which were constructed to satisfy the following conditions:

> a) the principle of contradiction, in the form $\neg(A \wedge \neg A)$, should not be valid in general;
>
> b) from two contradictory premises A and $\neg A$ we should not be able to deduce any formula whatsoever;
>
> c) the most important schemes and rules of classical logic compatible with a) and b) should be retained (thus this is essentially a conservative approach).

A description of these calculi can be found in da Costa (1974) and da Costa and Marconi (*op. cit.*). The above hierarchy can be extended to corresponding hierarchies of first-order predicate calculi, with or without equality, and to theories of descriptions, on which it is possible to found paraconsistent set theories (da Costa 1986).

The basis of this whole approach is to take a valuational semantics for the "positive" logic and then lay down the conditions on the evaluation of $\neg A$ anew. It can be shown that the above calculi have an interesting semantics of valuations relative to which they are sound and complete (da Costa and Alves 1977) and this work has lead to the development of a general theory of valuations which can be applied to any logic whatsoever (Loparic and da Costa 1984).

Since these logics involve non-standard rules for the logical constants, it can be objected that paraconsistent negation is not "genuine" or "natural" negation and should be rejected. However, it is not clear how this notion of natural negation can be characterised independently of its identification with classical negation. One possibility is to refer to "natural" linguistic usage; however, it is not at all clear that such usage unequivocally supports the classical conception (Marconi 1984). Certainly the "natural" form of negation cannot be argued for on the basis of a consideration of the occurrence or non-occurrence of "genuine" contradictions in our linguistic and behavioural experience, since the existence of such contradictions in certain situations is precisely what we are claiming here.

Elsewhere it has been argued that a logical connective should be regarded as 'natural' if it captures enough features, both syntactic and semantic, of the natural expressions which it is intended to represent (da Costa and Marconi, *op. cit.*). Taking classical logic as a standard, a connective can be considered a modified negation connective provided the changes with respect to the classical form are neither too big nor too arbitrary (da Costa 1982; this therefore marks off da Costa's approach from Priest's, since the latter explicitly espouses the overthrow of the classical edifice, Priest 1987). Given this, Smith's claim that ". . .da Costa's negation operator simply cannot be given (a) natural language interpretation." (*op. cit.*, p. 245), seems to be rather wide of the mark.

There is obviously a great deal more which can be, and, indeed, has been, said about these systems. Our principal concern here, however, is merely to delineate their main features, the most important of which, as far as the present paper is concerned, is the ability to accommodate strict inconsistency, such as occurs, we claim, in cases of self-deception. In (da Costa and French 1989a) we have outlined a paraconsistent logic of belief, there denoted S_1, which can accommodate certain forms of inconsistent or contradictory beliefs. In the appendix below, as we have already mentioned, we present a similar system which seems particularly suitable for capturing the inconsistent aspect of self-deception, as characterised above.

Conclusion

What is a paradox? A common view is that it can be understood as: "an apparently unacceptable conclusion derived by apparently acceptable reasoning from apparently acceptable premises." (Sainsbury 1988, p. 1). What, then, is unacceptable in the "paradox" of self-deception? Clearly the holding of contradictory or inconsistent beliefs. Why is this unacceptable? Because, the standard line runs, it violates the principle of contradiction. The next question is less frequently asked, however: why should we accept this principle? The answer may seem obvious if inconsistency is taken to be logically identical with triviality, but the development of paraconsistent logics shows this to be an unwarr

classical logic while weakening, or rejecting entirely, the above principle.

Thus the holding of strictly inconsistent beliefs is unacceptable only if we assume an underlying classical logical framework. Dispensing with this framework removes the sting from the paradox and allows us to accept it at face value. Within a paraconsistent belief system there is no threat of a collapse into doxastic triviality and we are free to explore the philosophical and psychological ramifications of the problem without being restricted in our investigations by the particular form of classical logic.

It seems clear that the above assumption is made by the majority of people working on the problem of self-deception. Although we hesitate to introduce Lakatosian or Kuhnian terms into the present context, given the acknowledged difficulties associated with all of the supposed resolutions currently on offer, perhaps it is time to switch "paradigms" or move over to an alternative "research programme" based on a different logical "hard core." At the very least, this possibility is worth exploring.

Indeed, it almost seems forced upon us when we consider the examples of "true" self-deception mentioned previously. In these cases discussion of willful ignorance, attention shifts etc., seems particularly inappropriate. This is brought out even more clearly when the problem is located within the context of inconsistency and contradiction in general, as we have tried to do above. Of course, once this possibility is accepted, the discussion is not over. There exist many (in fact infinitely many) paraconsistent systems and a choice must be made as regards which is the most appropriate for modelling this particular situation. In the preceding section we briefly set down the main characteristics of the three broad kinds of paraconsistent logic currently under consideration and below we present a particular system of paraconsistent doxastic logic within the "positive plus" approach of da Costa.

Finally, we repeat that formal techniques are not everything and that there is obviously a great deal more to be done with regard to this problem. However, recalling Haight's statement that ". . .the idea is self-contradictory and the thing therefore impossible.", we hope to have at least demonstrated the dangers involved in drawing conclusions about the existence of certain phenomena from premises embedded in a particular form of logic.[2]

University of São Paulo, Southeast Missouri State University

Received December 15, 1989

NOTES

1. See the brief summary in da Costa and French, forthcoming, ed.

2. We would like to thank the Institute of Advanced Study of the University of São Paulo and the Centre of Logic, Epistemology and History of Science of the University of Campinas for their generous support during the preparation of this work. One of us (French) would also like to acknowledge funding received from the Brazilian National Council for Scientific and Technological Development (CNPq).

APPENDIX

We now present a simple system of paraconsistent doxastic logic which does not exclude ab initio the possibility that a person may possess, for a certain time, contradictory beliefs. This system also does not rule out the possibility of permanent contradictory beliefs and logical contradictions. We denote our logic by $\mathcal{B}$ and note that it constitutes a paraconsistent doxastic propositional logic (for more details on this kind of logic, see da Costa and French 1989a).

Description and Principal Properties of $\mathcal{B}$:

Primitive symbols of $\mathcal{B}$: 1) $\rightarrow$ (implication); $\wedge$ (conjunction); $\vee$ (disjunction) ;$\neg$ (negation), and B (the doxastic operator); $\leftrightarrow$ (equivalence) is defined as in the classical propositional calculus; 2) propositional variables: a denumerably infinite set of propositional variables; 3) parentheses.

The concept of formula is defined as usual, and they will be represented by small Greek letters.

Postulates (axiom schemes and primitive rules) of $\mathcal{B}$:

$\rightarrow_1$ $\alpha \rightarrow (\beta \rightarrow \alpha)$ $\quad \rightarrow_2$ $(\alpha \rightarrow \beta) \rightarrow ((\alpha \rightarrow (\beta \rightarrow \gamma) \rightarrow (\alpha \rightarrow \gamma)$ $\quad \rightarrow_3$ $\dfrac{\alpha \; \alpha \rightarrow}{\beta}$ $\quad \rightarrow_4$ $((\alpha \rightarrow \beta) \rightarrow \alpha) \rightarrow \alpha$

$\wedge_1$ $(\alpha \wedge \beta) \rightarrow \alpha$ $\quad \wedge_2$ $(\alpha \wedge \beta) \rightarrow \beta$ $\quad \wedge_3$ $\alpha \rightarrow (\beta \rightarrow (\alpha \wedge \beta))$

$\vee_1$ $\alpha \rightarrow (\alpha \vee \beta)$ $\quad \vee_2$ $\beta \rightarrow (\alpha \vee \beta)$ $\quad \vee_3$ $(\alpha \rightarrow \gamma) \rightarrow ((\beta \rightarrow \gamma) \rightarrow ((\alpha \vee \beta) \rightarrow \gamma))$

$\neg_1$ $\dfrac{\alpha \leftrightarrow \beta}{\neg\alpha \leftrightarrow \neg\beta}$ $\quad \neg_2$ $\alpha \vee \neg\alpha$ $\quad \neg_3$ $\neg\neg\alpha \rightarrow \alpha$

B_1 $B(\alpha \rightarrow \beta) \rightarrow (B\alpha \rightarrow B\beta)$ $\quad B_2$ $B(\alpha \wedge \beta) \leftrightarrow (B\alpha \wedge B\beta)$ $\quad B_3$ $(B\alpha \vee B\beta) \rightarrow B(\alpha \vee \beta)$

B_4 $\dfrac{\alpha}{B\alpha}$ $\quad B_5$ $B\alpha \leftrightarrow BB\alpha$

These postulates clearly possess an informal, intuitive motivation, although we shall not enter here into a discussion of the reasons underlying them. The following remarks are worth making, however: 1) Since $B\alpha$ means that a certain person x, usually kept implicit, believes that α is the case, it seems natural to accept B_1 to B_5. On the other hand, since we are constructing a logic that does not rule out contradictory beliefs, the common postulate $B\alpha \rightarrow \neg B\neg\alpha$ (da Costa and French *ibid.*) of classical doxastic logic apparently must be considered invalid from our point of view (of course, it could always be conjoined to the above list with little difficulty); 2) $\neg\neg\alpha \rightarrow \alpha$ is a valid scheme of $\mathcal{B}$ but $\alpha \rightarrow \neg\neg\alpha$ is not, for the following reason: since our logic does not exclude the possibility that α and $\neg\alpha$ are both true (i.e. it does not exclude the possibility of true logical contradictions) we could have that α is true while $\neg\neg\alpha$ is false, which runs contrary to our implication above (which is essentially the classical implication by postulates $\rightarrow_1$ to $\rightarrow_4$).

We define the concepts of theorem, of syntactic consequence ($\vdash$) etc., as usually given.

Theorem 1: In $\mathcal{B}$ all valid formulas and rules of the classical propositional calculus which do not involve negation are also valid.

Proof: Postulates $\rightarrow_1$ to $\vee_3$ show that $\mathcal{B}$ contains the classical positive propositional logic.

Theorem 2: In $\mathcal{B}$ we have:

$\nvdash (\alpha \wedge \neg\alpha) \rightarrow \beta$ $\quad \nvdash (\alpha \wedge \neg\alpha) \rightarrow \neg\beta$ $\quad \nvdash (\alpha \leftrightarrow \neg\alpha) \rightarrow \beta$

$\nvdash (\alpha \leftrightarrow \neg\alpha) \rightarrow \neg\beta$ $\quad \nvdash \neg(\alpha \wedge \neg\alpha)$ $\quad \nvdash \alpha \leftrightarrow \neg\neg\alpha$

$\nvdash (B\alpha \wedge B\neg\alpha) \rightarrow \beta$ $\quad \nvdash (B\alpha \wedge B\neg\alpha) \rightarrow B\beta$

$\nvdash \alpha \rightarrow \neg B\neg\alpha$

Proof: By the use of convenient logical matrices.

$\mathcal{B}$ has a smooth semantics of valuations relative to which it is sound and complete. (cf. da Costa and French *ibid.*) We can define, without difficulty, the notions of semantic consequence, of a model of a set of formulas, of consistent set of formulas, of inconsistent set of formulas, etc.

Theorem 3: In $\mathcal{B}$ there are inconsistent sets of formulas which have models.

Proof: By means of the semantics of valuations of $\mathcal{B}$ (da Costa and Carnielli 1987).

$\mathcal{B}$ can easily be extended to first-order and higher order logics. On the other hand, employing other paraconsistent principles (for instance, those corresponding to da Costa's hierarchy Cn, $1 \leq n \leq w$, (see Arruda 1980) we can formalise paraconsistent doxastic logics which are different than $\mathcal{B}$. These calculi seem to reflect important characteristics of actual belief systems, including those relating to self-deception.

In particular, and given what we have said in the main body of this paper, in order to transform $\mathcal{B}$ into a logic *for* self-deception, it is convenient to add to it the following postulate:

B_6 $B\neg\alpha \rightarrow \neg B\alpha$

This, of course, does not modify $\mathcal{B}$'s main characteristics.

We re-emphasise that we are not trying to construct another kind of logic with the aim of replacing the classical form in those situations in which it has been successfully used. Our intention is much more modest: we are simply arguing that if one wants to systematize those issues or situations which involve inconsistent or contradictory beliefs, it is perhaps convenient and worthwhile to take advantage of the extant systems of paraconsistent logic. To put it even more explicitly, we are proposing the use of more general systems of logic to cope with doxastically inconsistent bodies of belief, as in the case of self-deception.

We feel that the general discussion above is sufficient to make clear the nature of a logic compatible with true logical contradictions and with contradictory beliefs. Of course, it is also possible to construct logical systems which rule out the former but not the latter (da Costa and French *op. cit.*).

BIBLIOGRAPHY

Arruda, A.I. (1980). "A Survey of Paraconsistent Logic," in A.I. Arruda, R. Chuaqui and N.C.A. da Costa, eds., *Mathematical Logic in Latin America* (North Holland, 1980), pp. 1-41.

_____(1989). "Aspects of the Development of Paraconsistent Logic," in Priest, Routley and Norman (1989).

Bok, S. (1980) "The Self-Deceived," *Social Science Information*, vol. 19, pp. 923-35.

Cherniak, C. (1984) "Computational Complexity and the Universal Acceptance of Logic," *The Journal of Philosophy*, vol. 81, pp. 739-58.

da Costa, N.C.A. (1974) "On The Theory of Inconsistent Formal Systems," *Notre Dame Journal of Formal Logic*, vol. 11, pp. 497-510.

da Costa, N.C.A. (1982) "The Philosophical Import of Paraconsistent Logic," *The Journal of Non-Classical Logic*, vol. 1, pp. 1-19.

da Costa, N.C.A. (1986). "On Paraconsistent Set Theory," *Logique et Analyse*, vol. 115, pp. 361-371.

da Costa, N.C.A. and Alves. E.H. (1977). "A Semantical Analysis of the Calculi Cn,"*Notre Dame Journal of Formal Logic*, vol. 18, pp.621-30.

da Costa, N.C.A. and Carnielli, W.A. (1986) "On Paraconsistent Deontic Logic," *Philosophia,* vol.6, pp. 293-303.

da Costa, N.C.A. and French, S. (1988) "Pragmatic Probability, Logical Omniscience and the Popper-Miller Argument," *Fundamenta Scientae,* vol. 9, pp. 43-53.

da Costa, N.C.A. and French, S. (1989a) "On the Logic of Belief," *Philosophy and Phenomenological Research*, vol. 49, pp. 431-46.

da Costa, N.C.A. and French, S. (1989b) "Pragmatic Truth and the Logic of Induction," *British Journal for the Philosophy of Science*, vol. 40, pp. 333-56.

da Costa, N.C.A. and French, S. (forthcoming a) "The Model Theoretic Approach in Philosophy to Science," to appear in *Philosophy of Science*.

da Costa, N.C.A. and French, S. (forthcoming b) "Relativism, Models and Representational Belief," in preparation.

da Costa, N.C.A. and French, S. (forthcoming c) "Review of G. Priest, *In Contradiction*," to appear in *History and Philosophy of Logic*.

da Costa, N.C.A. and French, S. (forthcoming d) "Ontology and Paraconsistency," to appear in H. Burkhardt and B. Smith, eds., *The Handbook of Metaphysics and Ontology*,. (Munich: Philosophia Verlag).

da Costa, N.C.A. and Marconi, D. (1989) "An Overview of Paraconsistent Logic in the '80s," *Logica Nova*, (Berlin: Akademie-Verlag).

Demos, R. (1960). "Lying to Oneself," *The Journal of Philosophy*, vol. 57, pp. 588-95.

Fingarette, H. (1969) *Self-Deception* (London: Routledge and Kegan Paul, 1969).

French, S. (forthcoming). "Rationality, Consistency, and Truth," Read at the A.P.A. Central Division Meeting, Chicago April 27-29, 1989 (forthcoming in *The Journal of Non-Classical Logic*).

Garber, D. (1983) "Old Evidence and Logical Omniscience in Bayesian Confirmation Theory," in J. Earman (ed.), *Testing Scientific Theories* (Minneapolis: University of Minnesota Press, 1983), pp. 99-132.

Glick, I.D., Wiess, R.S. and Parkes, C.M. (1974) *The First Year of Bereavement* (Witley, 1974).

Graham, G. (1986) "Russell's Deceptive Desires," *Philosophical Quarterly*, vol. 36, pp. 223-29.

Haight, M. (1980) *A Study of Self Deception* (Sussex: Harvester Press, 1980).

_____(1985) "Review of D. Pears, *Motivated Irrationality*," *Philosophical Books*, vol. 26, pp. 48-50.

Harman, G. (1986) *Change in View* (Cambridge, MA: MIT Press, 1986).

Hintikka, J. (1962) *Knowledge and Belief* (Ithaca: Cornell University Press, 1962).

Kyburg, H. (1961) *Probability and the Logic of Rational Belief* (Middletown, CT: Wesleyan University Press, 1961).

_____(1987) "The Hobgoblin," *The Monist*, vol. 70, pp. 141-51.

Loparic, A. and da Costa, N.C.A. (1984) "Paraconsistency, Paracompleteness and Valuations," *Logique et Analyse*, vol. 106, pp. 119-31.

Marconi, D. (1981) "Types of Non-Scotian Logic," *Logique et Analyse*, Vol. 95-96, pp. 407-14.

(1984) "Wittgenstein on Contradiction and the Philosophy of Paraconsistent Logic," *History of Philosophy Quarterly*, vol. 1, pp. 333-52.

Martin, M.W. (1986) *Self Deception and Morality* (Lawrence, Kansas: University Press of Kansas, 1986).

McLaughlin, B. and Okensberg-Rorty, A. (eds.) (1988) *Perspectives on Self-Deception* (Berkeley: University of California Press, 1988).

Mele, A. (1983) "Self-Deception," *Philosophical Quarterly*, vol. 33, pp. 365-377.

Mele, A. (1987) "Recent Work on Self-Deception," *American Philosophical Quarterly* , vol. 24, pp. 2-17.

Mikenberg, I., Chuaqui, R. and da Costa, N.C.A. (1986) "Pragmatic Truth and Approximation to Truth," *Journal of Symbolic Logic*, vol. 51, pp. 201-21.

Moriconi, E. (1981) "Logica e Contaddizione," *Theoria*, vol. 2, pp. 83-130.

Norton, J. (1987) "The Logical Inconsistency of the Old Quantum Theory of Black Body Radiation," *Philosophy of Science*, vol. 54, pp. 327-50.

Pereira, A. and French, S. (forthcoming) "Metaphysics, Pragmatic Truth and the Underdetermination of Theories," to appear in *Dialogos*.

Priest, G. (1987) *In Contradiction* (Dordrecht: Martinus Nijhoff, 1987).

Priest, G. and Routley, R. (1984) "Introduction: Paraconsistent Logics," *Studia Logica*, vol. 43, pp. 3-16.

Priest, G., Routley, R. and Norman, J. (eds.), (1989) *Paraconsistent Logic : Essays on the Inconsistent* (Munich: Philosophia Verlag, 1989).

Pugmire, D. (1969) "Strong Self-Deception," *Inquiry*, vol. 12, pp. 339-46.

Rescher, N. (1968) *Topics in Philosophical Logic* (Dordrecht: Reidel, 1968).

Rescher, N. and Brandom, R. (1980) *The Logic of Inconsistency* (Oxford: Blackwell, 1980).

Rorty, A. (1972) "Belief and Self-Deception," *Inquiry*, vol. 5, pp. 387-410.

Routley, R., Meyer, R.K., Plumwood, V. and Brady, R. (1982) *Relevant Logics and Their Rivals* (Atascadero, CA: Ridgeview, 1982).

Sainsbury, R.M. (1988) *Paradoxes* (Cambridge: Cambridge University Press, 1988).

Schotch, P.K. and Jennings, R.E. (1980) "Inference and Necessity," *Journal of Philosophical Logic*, vol. 9, pp. 329-40.

Smith, J. (1988a) "Scientific Reasoning or Damage Control: Alternative Proposals for Reasoning with Inconsistent Representations of the World," *PSA 1988*, vol. 1, Philosophy of Science Association, pp. 241-48.

_____(1988b) "Inconsistency and Scientific Reasoning," *Studies in History and Philosophy of Science*, vol. 19, pp. 429-45

Sperber, D. (1982) "Apparently Irrational Beliefs," in M. Hollis and S. Lukes (eds.), *Rationality and Relativism* (Cambridge, MA: MIT Press,1982).

Szabados, B. (1974) "Rorty on Belief and Self-Deception," *Inquiry*, Vol. 17, pp. 464-73.

Ullmann, L.P. and Krasner, L. (1969) *A Psychological Approach to Abnormal Behavior* (Englewood Cliffs:Prentice Hall, 1969).

van Fraassen, B. (1985) "Empiricism in the Philosophy of Science," in P. Churchland and C. Hooker (eds.), *Images of Science* (Chicago: University of Chicago Press, 1985).

Waters, C.K. (1987) "Relevance Logic Brings Hope to Hypothetico-Deductivism," *Philosophy of Science*, vol. 54, pp. 453-64.

Williams, J. (1982) "Believing the Self-Contradictory," *American Philosophical Quarterly*, vol. 19, pp. 279-85.

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 27, Number 3, July 1990

RECENT OBITUARIES OF EPISTEMOLOGY

Susan Haack

A traditional conception takes the central tasks of epistemology to be the *explication of epistemic concepts*, central among which, on the plausible assumption that justified true belief is at least a necessary condition of knowledge, is the concept of epistemic justification; and the *ratification of criteria of justification*, which, on the plausible assumption that truth is at least an essential component of the goal of inquiry, centrally involves investigating the relation between justification and truth. And the traditional approach conceives of these projects as distinctively philosophical in a strong sense: they are to be undertaken, not by any kind of empirical investigation, but a priori. It seems appropriate to call this conception "traditionalist apriorism."

Traditionalist apriorism still has its defenders;[1] but it also has rivals. Reformists accept the legitimacy of the traditional epistemological projects, but repudiate the apriorist approach, proposing, instead, to undertake them in a naturalistic fashion. Revolutionaries repudiate the traditional projects altogether, holding them misconceived.

Reformists are naturalists, but not all naturalists are reformists. In one sense of that seductive but, or perhaps because, ambiguous phrase, "epistemology naturalized" refers to the idea that the traditional epistemological projects should be undertaken, not a priori, but within the web of empirical belief ("modest reformist naturalism"); in another, to the more ambitious idea that the traditional epistemological projects should be undertaken within science ("scientistic reformist naturalism"); in another, to the radical idea that the traditional projects are misconceived, and should be repudiated in favour of scientific investigations of human cognitive processing ("revolutionary naturalism"). A further complication is that some who describe themselves as doing "naturalistic epistemology" seem simply to be extending the term "epistemology" to apply to scientific as well as philosophical investigations of cognition, without prejudice to the legitimacy either of the traditional epistemological projects or of the apriorist approach to those projects ("expansionist naturalism").[2]

Some naturalists are revolutionaries, but not all revolutionaries are naturalists. "Revolutionary nihilism," as it seems appropriate to call it, like revolutionary naturalism, urges the repudiation of the traditional projects, but, unlike revolutionary naturalism, instead of proposing some scientific successor-subject, denies that any successor-subject is either necessary or desirable.[3]

Of course, there is a degree of vagueness about what should count as pursuing reformulated but recognisably continuous versions of the traditional projects, and what as pursuing new projects altogether, and hence about the distinction between reformists and revolutionaries. Of course, also, there is a question about what, if the traditional projects are misconceived, should count as a successor-subject to epistemology, and what as simply a new occupation for ex-epistemologists, and hence about the distinction between revolutionary naturalists and revolutionary nihilists. My purpose here, however, is only to defend the legitimacy of the traditional projects against the arguments of the revolutionaries; and for this purpose these complications may safely be ignored.

Epistemology has been pronounced dead before;[4] but of late there seems to be a serious danger that the ever more persistent rumors of its demise will be believed. The problems of epistemology are formidably difficult; and the work done in epistemology over the last century or so has, perhaps, done as much to point up how intractable they are as to resolve them. But my view of the matter is resolutely optimistic: elsewhere I have argued that a very modest reformist naturalism can resolve at least some of the difficulties;[5] and here I shall argue that no good

reason has been given for thinking the traditional projects misconceived.

Quine's seminal paper, "Epistemology Naturalized," seems to me ambivalent between a reformist and a revolutionary naturalism;[6] and disentangling the revolutionary from the reformist strands is not a task I can undertake here. My present concern is with some more recent, and more unambiguous, revolutionaries: Rorty, Stich and the Churchlands.[7] Though Rorty qualifies as a revolutionary nihilist and Stich and the Churchlands, rather, as revolutionary naturalists, it is worthy of note that both Rorty and Patricia Churchland derive inspiration from Kuhn, urging the overthrow of the old epistemological paradigm; and that all four of my revolutionaries sound a note of ambitious wistfulness for greener pastures than the old, overgrazed epistemological fields.[8] But Rorty's critique of the traditional epistemological questions is quite different from the others', so I shall take him separately.

I

According to Rorty, the traditional conception is that the role of epistemology is to supply the foundations for science, to legitimate the claims of science to give us genuine knowledge. But this, like the conception of the role of philosophy in which it is embedded, depends on the idea that there is a sharp distinction between science and philosophy, or more specifically between science and the philosophical theory of knowledge. And this distinction is quite recent; it was not acknowledged by Hobbes, or even by Descartes, but has come to seem obvious only since Kant—and, like the conception of epistemology it encourages, it is "optional." The philosophical theory of knowledge happens to have developed, furthermore, under the influence of an optical metaphor, an analogy between knowing and seeing, which has encouraged the idea that the relation between conceptual and sensory elements in knowing is fundamental, and has seemed to legitimise the ambition to establish the *bona fides* of scientific knowledge a priori, by appeal to the formal characteristics of the human mind, the "organ" of knowing. But this optical metaphor, like the epistemological theories it encourages, is "optional."[9]

And this "foundationalism," as Rorty calls it, is misconceived. Sellars' critique of the notion of the given, and Quine's of the analytic and hence, by implication, of the a priori, are, together, sufficient to show it untenable. The idea that the justification of a belief lies in its relation, direct or indirect, to what is given in experience, is a misconception, a misconception resulting from a confusion, encouraged by the optical metaphor, of justification with causation.[10]

Justification, according to Rorty, is better conceived, rather, as a conversational matter; for a belief to be justified is for it to be defensible against the objections of one's epistemic community. And the differing criteria of different communities are "incommensurable"; no agreement can be expected as to what standards of defending beliefs are best. Nor does it make sense to seek to ratify these or those criteria of justification by arguing that beliefs which satisfy them are likely to be true; for this requires the idea of truth as correspondence, as faithful picturing—another legacy of the optical metaphor, and covertly unintelligible. Justification is not only a conversational, but also a purely conventional, matter; our practices of defending and criticizing beliefs cannot be grounded in anything external to those practices. Epistemology is misconceived; conceived, one might say, in the sins of confusing justification with causation, the conventional with the objective.[11]

Rorty urges, not the replacement of epistemology by some natural-scientific successor-subject, but the repudiation, as it were, of the idea that the rejection of epistemology leaves any gap that needs to be filled. He thinks there remains a role for the ex-epistemologist; but it is "hermeneutic" rather than epistemological, "edifying" rather than systematic, a matter of "carrying on the conversation" rather than of trying to commensurate incommensurable discourses.[12]

Happily, it is not necessary to engage in detailed discussion of Rorty's claims about the history of epistemology; for even if it were true that the problems now thought of as central to epistemology have come to be perceived as problems only since Kant, it wouldn't follow that those issues are misconceived or not genuinely problematic. Nor is it necessary to engage in detailed consideration of his claims about the influence of the optical metaphor; for even if it were true that the problems of epistemology have come to be approached as they now are because of the dominance of this metaphor, it wouldn't follow that those approaches are wrong or fruitless.[13]

The focus must be on Rorty's arguments that what he calls "foundationalism" is not just optional, but misconceived. But an initial strategic difficulty is that his arguments do not all have the same target; for Rorty uses the term "foundationalism" to refer to at least three distinct positions. I will make the necessary distinctions as follows:

> "[Empiricist] foundationalism" will refer to any theory of justification which distinguishes basic beliefs, held to be justified, independently of the support of other beliefs, by the subject's experience or their relation to the external world, and derived beliefs, held to be justified by the support, direct or indirect, of basic beliefs; i.e., which postulates basic beliefs as the foundations of knowledge.
>
> "*foundationalism*" will refer to any approach combining a strong conception of knowledge requiring true belief justified by criteria which not only are, but have been shown to be, satisfactory, with an apriorist approach to the project of ratification; i.e., which regards the a priori as the foundation of epistemology.
>
> "FOUNDATIONALISM" will refer to the thesis that criteria of justification are not purely conventional but objectively grounded, being satisfactory only if related in an appropriate way to the goal of inquiry; i.e., which takes criteria of justification to be founded by their relation to the goal of inquiry.

FOUNDATIONALISM does not imply *foundationalism*, nor *foundationalism* foundationalism. It could be that criteria of justification are satisfactory only if appropriately related to the goal of inquiry, but that the way to show that criteria are or are not satisfactory is, not a priori, but within empirical knowledge; then FOUNDATIONALISM would be true but *foundationalism* false. It could be that the way to show that criteria of justification are or are not satisfactory is a priori, but that the satisfactory criteria are, not foundationalist, but coherentist; then *foundationalism* would be true but foundationalism false.

The allegation of a confusion of justification with causation, like the appeal to Sellars' critique of the given, is relevant to the truth of foundationalism; the appeal to Quine's critique of the analytic, to the truth of *foundationalism*; and the repudiation of the correspondence theory, to the truth of FOUNDATIONALISM. The other arguments merit attention as a way of getting clearer about the alternatives to foundationalism and *foundationalism*; but it should be clear already that it is on FOUNDATIONALISM, not foundationalism or *foundationalism*, that the legitimacy of epistemology depends, and hence that it is on the argument about truth that Rorty's case against epistemology rests.

It is possible that some foundationalists of an externalist stripe have hoped that an account of the etiology of a belief would be sufficient to show that it is, or that it is not, justified.[14] If so, they are mistaken; justification cannot be identified with causation. And foundationalists of an experientialist stripe, who hold that basic beliefs are justified by the support of a subject's experience, have sometimes talked, less than helpfully, of beliefs' "confronting experience," or of "the information conveyed by the senses," or slipped conveniently from locutions like "A sees an F" to locutions like "A sees that x is an F." But it is false that foundationalism inevitably involves a confusion of justification with causation. The truth is, rather, that the concept of justification is a "causical" concept, i.e., a concept which combines causal and logical elements. Whether, or to what degree, a subject is justified in some belief of his depends, not just on *what* he believes, but also on *why* he believes it. To explicate the concept of justification fully, therefore, it would be necessary to distinguish the state from the content senses of "belief" and to articulate the interaction between the causal and the logical aspects of justification.[15] Causical foundationalism is innocent of the charge of confusion.

Nor is Sellars' critique of the given[16] sufficient to preclude all forms of foundationalism. It threatens those strong forms according to which basic beliefs are fully or conclusively justified by the support of experience; it is debatable, however, that it has the same force against those weaker versions according to which basic beliefs are justified to some degree, or *prima facie*, by experience.

However, though Rorty's arguments fall well short of establishing this, it is, I think, true that foundationalism, even a weak, causical, experientialist foundationalism, fails. For weak foundationalism faces a difficulty that Rorty doesn't mention: in virtue of its insistence on the one-directional character of relations of support, it is powerless to explain how it could be that the supposedly basic beliefs which are *prima facie* justified by experience could come to be *more*, or *less*, justified in

the light of further beliefs, e.g., about the conditions of perception.

But it doesn't follow from the fact that foundationalism fails that one is obliged to accept something along the lines of the conversationalist account of justification that Rorty proposes. These are obviously not the only alternatives. One might, for instance, like Davidson, who claims as Rorty does that foundationalism rests on a confusion of justification with causation, opt for some form of coherentism.[17] Or one might, like me, opt for a causical foundherentism, according to which, though experience is relevant to justification, there need be no privileged class of beliefs justified, independently of the support of other beliefs, by experience alone.[18]

Rorty's argument against *foundationalism* is also inconclusive. Even if it is true that Quine has shown the notion of analyticity to be indefensible,[19] it would follow that no a priori ratification of epistemic criteria is possible only on the assumption that only analytic truths can be known a priori.

Once again, however, though Rorty's argument is inconclusive, his conclusion is, I believe, true. Ironically, though, the most persuasive argument for this conclusion needn't depend on the repudiation of analyticity at all; better, I should have thought, to acknowledge that the ratification of criteria of empirical justification is bound to require synthetic assumptions, and to rely on a repudiation of the synthetic a priori.[20]

But it doesn't follow from the fact that apriorism, and hence *foundationalism*, fails, that the project of ratification is misconceived. These are obviously not the only alternatives. If an a priori ratification is impossible, perhaps an a posteriori ratification is possible, as reformist naturalists hope. And even if no ratification of criteria of justification is possible, as those pessimists fear who see no hope of avoiding vicious circularity, it wouldn't follow that the question of ratification is senseless, that criteria of justification are merely conventional.

Given the significance Rorty attaches to the fact that the idea of a sharp distinction between science and philosophy is quite recent, it may be that he has in mind some such argument as this: once the idea is abandoned that philosophy deals with the sphere of the a priori, science with the a posteriori, the idea of a *distinctively philosophical* theory of knowledge is seen to be untenable. But if this is what he is thinking, it misses a significant subtlety—though some of the blame perhaps falls on Quine, who uses the term "science" sometimes narrowly, to refer to the natural sciences specifically, and sometimes broadly, to refer to our empirical knowledge generally. Giving up the idea that philosophy is distinguished from science by its a priori character gives rise to a picture—with which, like Quine, I sympathise—of philosophy as continuous with natural science. But it in no way obliges one to deny that there is a difference *of degree* between philosophy and the natural sciences; it remains plausible to think that those problems qualify as belonging to philosophy which are most general and abstract, while those problems qualify as belongings to the natural sciences which are more specific and concrete. So it by no means follows that any legitimate question about knowledge must be answerable by the natural sciences. It is, of course, at just this point that the modest reformist naturalist parts company with the scientistic naturalist.

Rorty is critical of Quine's attempt to psychologize epistemology; and he is correct, I believe, in thinking that there is no realistic prospect that the traditional epistemological questions could be answered within psychology. But only if it were established that any legitimate question about knowledge must be answerable by the natural sciences would this have any tendency to support the conclusion that the traditional epistemological questions are illegitimate.

So the weight of Rorty's case against epistemology rests on considerations about truth. Now, if it were true that the notion of epistemic criteria being satisfactory or unsatisfactory can be given sense only by reference to a transcendental, correspondence, "mirroring" conception of truth, and that such a conception is unintelligible, it would undeniably follow that the project of ratification is misconceived; and this would indeed undermine the whole enterprise of epistemology as it has traditionally—I would say, "since Descartes," but am willing, for the sake of argument, to settle for "since Kant"—been conceived.

However, though the argument is valid, it has at least one false premiss. On my understanding the intelligibility of the notion of these or those criteria of justification being satisfactory or unsatisfactory *does* depend on the notion of truth; for criteria of justification to be satisfactory is for them to be appropriately related to the goal of inquiry, and the

goal of inquiry, according to me, is *as complete as possible a true, explanatory account of how things are*. But so far as I can tell the only demand placed on the concept of truth by this role is that it be what I call "minimally realist." Let me explain. By a "radical relativist" theory of truth I mean one according to which the predicate "true" makes sense only relative to some community, individual or theory, as in: "true-in-T," "true-for-A," etc. Any theory which denies this, I count as minimally realist. Both a pragmatist conception, according to which S is true just in case it belongs to the hypothetical ideal theory, and a transcendentalist conception, which allows the possibility that even the best theory possible for us may be false or incomplete, qualify as minimally realist. I see no reason to suppose that the problem of ratification makes sense only from a transcendentalist perspective.

I suspect that Rorty runs together minimal realism and transcendentalism because he is misled by the fact that "correspondence" can be used to refer to either of two sorts of theory of truth: what I think of as "correspondence without teeth," by which I mean, simply, an acknowledgement of the formula "a belief is true just in case it corresponds to the facts" which interprets it as just an idiomatic way of saying what might also be said by "a belief is true just in case things are as it says," or, "for any *p*, it is true that *p* just in case really, *in fact*, *p*," where "in fact" shows its true colours as an emphatic adverb; and "correspondence with teeth," which not only accepts the correspondence formula, but also interprets it literally, as requiring the postulation of facts as theoretical entities and correspondence as a theoretical relation. Correspondence without teeth amounts to little more than minimal realism. But it is correspondence with teeth that Rorty thinks unintelligible. Even if he were right, however, FOUNDATIONALISM does not depend on correspondence with teeth.[21]

This should not be taken as conceding that correspondence with teeth *is* unintelligible. The available candidates in this *genre*—Wittgenstein's and Russell's Logical Atomist theories, Austin's linguistic-conventional version—are, indeed, unsatisfactory; but it is unproven that there is anything necessarily incoherent about correspondence with teeth. Rorty seems to suspect that correspondence with teeth makes truth necessarily inaccessible to us, perhaps because correspondence with teeth is often described as defining truth in terms of a relation to "mind-independent facts." "Mind-independent fact," however, doesn't mean "fact in principle inaccessible to the mind" but "fact not essentially dependent on the mind," so the fear of inevitable inaccessibility is ill-founded. But the legitimacy of epistemology does not in any case depend on the intelligibility of correspondence with teeth.[22]

I have aspired here only to show that Rorty's arguments that epistemology is misconceived, fail, and not to attempt any defensive argument to establish the *bona fides* of the traditional problems. I don't want to leave consideration of Rorty, however, without commenting briefly on what he says about the role of the ex-epistemologist; for this manifests, I think, an ambivalence symptomatic of a half-suspected incoherence. One is entitled to ask why, in any case, one should expect there to *be* any role for the ex-epistemologist. Why, given that the project of alchemy was misconceived, would one expect there to be any role for the ex-alchemist? Certainly Rorty's explanations of what the ex-epistemologist is supposed to do are more baffling than edifying. At one point, one is told that he is to compare and contrast the incommensurable discourses which, as epistemologist, he confusedly hoped to commensurate; does this mean, one asks oneself, that he is to turn, after all, sociologist of knowledge? At another point, one is told that the ex-epistemologist is to study "abnormal" discourses; what could an "abnormal" discourse *be*, one asks oneself?—if an attempted conversation between participants from incommensurable discourses, what more illuminating conclusion could he hope to reach than that there is insuperable misunderstanding? And at yet another point one is told that the ex-epistemologist is to "carry on the conversation" of western culture; but what, one asks oneself, if the various discourses which constitute western culture really are incommensurable, could this be but participation in what he already knows must inevitably be misunderstanding?[23]

Rorty takes himself, presumably, to be doing what he thinks the ex-epistemologist is to do. Perhaps this is why his *modus operandi* seems so odd. He does not deny, of course, that we have, as a matter of social or conversational practice, criteria of what counts as good reasons, as flimsy evidence, as jumping to conclusions, and so on. Presumably he is engaging in those practices when he offers reasons for thinking that those practices are no more than conventional, that they could not be objectively

grounded by virtue of the likelihood that defensible beliefs are true. But if one really believed that our criteria of justification are purely conventional, are wholly without objective grounding, though one might conform to the practice, one would surely be obliged to adopt an attitude of cynicism towards it, to think of justification, as it were, in covert quotation marks. It is not that, in general, one cannot rationally engage in a practice one regards as wholly conventional; but that one cannot rationally fully, i.e., non-cynically, engage in a practice *of justifying beliefs* which one regards as wholly conventional. For *to believe that p is to believe that p is true*; so for one who denies that it even makes sense to suppose that there is any connection between a belief's being justified according to our practices, and its being true, it is impossible to see why a belief's being justified should be thought to have any bearing on whether one should hold it. No wonder Rorty looks so much as if he is sawing through the branch on which he is sitting.

II

Rorty's critique of epistemology is focussed on the question of the objectivity of criteria of epistemic evaluation; Stich's and the Churchlands' is more radical yet. Epistemology is centrally concerned with questions about the justification of belief; but the *bona fides* of the notion of belief are, they claim, threatened by developments in cognitive science. And if there are no such things as beliefs, then, as Stich puts it, the question, what beliefs we ought to hold, is as misconceived as the question, what deities we ought to propitiate.[24]

In answering Rorty's critique, the major problem was resolving ambiguities to determine which of his arguments supports what conclusion. In answering this second revolutionary critique, a significant initial hurdle is to disentangle the argument from the rhetoric.

Patricia Churchland's revolutionary piece, "Epistemology in the Age of Neuroscience," contains, so far as I can tell, no argument whatsoever against traditional epistemology; but its rhetoric is noteworthy. Epistemology, the message is, is out-of-date, is being superseded by developments in neuroscience. She phrases her announcement thus: "We are in the midst of a paradigm shift." The significance of the Kuhnian vocabulary soon becomes apparent. Churchland concedes that the old epistemological paradigm "has not been decisively refuted."[25] The concession is well-designed to convey without argument the impression that the old paradigm, if not *decisively* refuted, at least faces anomalies which pose a serious threat to its legitimacy; at the same time, the fact that in the Kuhnian picture paradigm-shifts are supposed to be more a matter of conversion than of rational argument operates in a subterranean way to seem to legitimise her failure to produce any arguments at all why the old paradigm is misconceived. It is impossible to avoid the suspicion that Churchland is urging conversion to the new, cognitive-scientific paradigm *simply on the grounds that it is the coming thing*. (There is a temptation to respond in rhetorical kind by dubbing her position "ambitious bandwagonism"; but I shall resist it!) The dizzying prospect she holds out of an epistemology revolutionised by cheap computing[26] should not be allowed to disguise the fact that no argument has been given for supposing revolution necessary.

Stich is a bit subtler. His title—*From Folk Psychology to Cognitive Science*—is a small masterpiece of suggestion. "Folk psychology" sounds as if it is bound to be crude, primitive, out-of-date; "cognitive science," in view of the favourable connotations of both "cognitive" and "science," sounds as if it is bound to be sophisticated, rigorous, up-to-date. And, as will shortly become clear, the suggestion carried by his title manages to insinuate itself into Stich's arguments against "folk psychology."

"Folk psychology," as Stich uses the phrase, refers to the idea—he likes to call it the "theory"—that human action is to be explained by reference to the agent's beliefs and or desires. The traditional problems of epistemology arise within, and presuppose, this theory. However, he argues, recent developments in cognitive science indicate that it is a serious possibility that folk psychology is simply a *false* theory, that its ontology is *mythical*, that there *are no* beliefs or desires.

Popper, anxious to avoid what he sees as the pitfalls of "psychologism," and to stress that scientists' attitude to their theories is, or should be, tentative, conjectural, noncommittal, inveighs against "belief philosophies" and urges the desirability of an "epistemology without a knowing subject" focussed exclusively on propositions and their logical relations. I have argued elsewhere that this kind of approach is

inadequate in principle.[27] Quine, anxious to remain within the constraints of behaviorism, and viewing the opacity of belief sentences with distaste, is inclined to discuss epistemological issues in terms of statements and subjects' dispositions to assent to/dissent from statements.[28] I sympathise with the hope of conducting epistemology within the confines of a concept of belief behaviorally constrained; I doubt that the hope of conducting it without even what one might call "behavioral-beliefs"—which would amount, in effect, to supposing reference to *dispositions* to behave eliminable—is realistic. For present purposes, anyway, I shall not challenge Stich's claim that the repudiation of beliefs would threaten the legitimacy of epistemology. My concern will be, specifically, with Stich's arguments that developments in cognitive science favour the repudiation of beliefs.

Stich offers the general argument that "however wonderful and imaginative folk theorising and speculation has been, it has turned out to be screamingly false in every domain where we now have a reasonably sophisticated science."[29] He admits that this is at best a very weak induction; in fact, it is utterly unconvincing. One problem is Stich's failure to distinguish a theory's being false and its ontology's being non-existent.[30] The main problem, though, is the shiftiness of his use of the term "folk theory." Stich talks casually, though hardly idiomatically, of "folk astronomy," "folk physics," etc., giving the impression that any old but now discredited theory would get counted as a "folk theory." If the adjective "folk" is applied to a theory or idea in virtue of the fact that it used to be widely accepted by the lay public *but is now discredited*, however, it is unproven that the idea that human action can be explained by the interaction of subjects' beliefs and desires *is* a folk theory, and the induction doesn't get off the ground. If, on the other hand, the adjective "folk" is being used in a neutral way, to refer to ideas which have been part of commonsense for a long time, the other premiss, that folk theories have invariably turned out to be false in the light of sophisticated scientific theorising, is false, and so again the induction doesn't get off the ground. What is being suggested, though not quite explicitly claimed, is that folk psychology stands to cognitive science as, say, ancient Babylonian to modern astronomy. But the implicit comparison is surely misleading. It is surely less plausible to think of modern psychology than to think of modern astronomy as a mature science; for psychology still seems today, as throughout its relatively short history, notably prone to schools and schisms, fads and fashions. The high-tech character of the techniques of cognitive science might warrant regarding it as *sophisticated*; but its theoretical background is not obviously of the rigour and strength that would warrant regarding it as *mature*. And even supposing the status of modern psychology as mature science granted for the sake of argument, it would be hard to see how Stich could allow force to the induction, "folk theories have usually turned out to be false, the belief-desire theory is a folk theory, so the belief-desire theory will probably turn out to be false," without also allowing force to the induction, "mature scientific theories have usually turned out to be false, modern psychology is a mature science, so its theories will probably turn out to be false."

This first line of Stich's has close affinities with an accusation made by Paul Churchland: that folk psychology is a degenerating research program.[31] Earlier in *Scientific Realism and the Plasticity of Mind* Churchland had argued for the appropriateness of calling the belief-desire model of the explanation of action a "theory" on the grounds that one's beliefs and desires are not incorrigibly open to introspective observation.[32] So the word "theory" is playing a neat rhetorical trick; it is a long way from granting the fallibility of introspection to describing the belief-desire model as a "research program." No sooner is the belief-desire model elevated to the status of "research program," however, than it is demoted by the accusation, "degenerating." This seems to me like calling the postulation of physical objects a "degenerating research program" on the grounds, first, that it has been sustained for many centuries without major modification, and, second, that it is simple, coarse-grained and gerrymandered relative to the ontology of modern physics.[33]

What, then, of Stich's claim that specific developments in cognitive science indicate that it is a serious possibility that there are no such things as beliefs? On the face of it, it is surprising to be told that cognitive science poses a threat to the *bona fides* of belief; for, unlike behaviorist psychology, cognitive psychology does not scruple to posit internal mental mechanisms, states and processes. Stich's claim, however, is that though cognitive psychology allows the postulation of internal states, it is open to

question whether, in particular, it calls for the postulation of beliefs and desires. He does not claim, and neither, so far as I am aware, is it true, that all or most or even much recent work in cognitive science is inhospitable to beliefs. But he argues that some recent work indicates that there may be no single class of mental states underlying and explaining, as beliefs are supposed to do, both verbal and non-verbal behavior; and that other work indicates that the isolatable units of cognitive processing are not, as beliefs are supposed to be, sentence-sized chunks, but something quite different. I note that he makes an attempt to explain—certainly it isn't obvious—how the two kinds of work allegedly inhospitable to beliefs are supposed to fit together; but I shall concentrate on showing that Stich's interpretation of the work he thinks makes "the case against belief" is insupportable.

The first of Stich's specific arguments appeals to work by Nisbett and Wilson on the phenomenon called "attribution." The central idea of attribution theory, as Stich reports it, is that people sometimes explain their own behavior by appeal to rather crude theories, and that this attribution of causes itself has behavioral effects; typical experiments in this field lead a subject to make a mistaken inference about the cause of some behavior of his, and then to behave as if this mistaken inference were correct. Stich discusses an experiment in which two groups of insomniac patients were given placebo pills; one group was told that the pills would produce rapid heart rate, irregular breathing, etc., i.e., the symptoms of insomnia, the other that they would produce lowered heart rate, regular breathing, etc. Attribution theory predicts that the first group would take less time to get to sleep, since they would attribute any arousal symptoms to the pills, while the second group would take more time to get to sleep, since they would infer that, since their arousal symptoms persist despite their having taken pills that should relax them, their thoughts must be especially distribing. Both predictions, apparently, were borne out. Questioning subjects about what they thought caused them to take more/less time to go to sleep, however, Nisbett and Wilson found that none offered what attribution theory conjectures to be the correct explanation; arousal subjects, apparently, typically replied that they found it easier to get to sleep later in the week. On the basis of several such experiments, to explain the discrepancy between subjects' verbal accounts of their mental processes and the hypothesized true explanations of their responses, Wilson proposes a model which he describes as postulating two relatively independent cognitive systems, one, largely unconscious, to mediate non-verbal behavior, and the other, largely conscious, to explain and verbalize what happens in the unconscious system.[34]

Stich claims that since beliefs are supposed to play a role in the explanation both of verbal and of non-verbal behavior, neither of Wilson's two systems could be thought of as a system of beliefs. But this interpretation seems to be a gratuitous piece of axe-grinding. If the true explanation of the time taken by subjects to get to sleep is as attribution theory says, there is indeed a discrepancy between the true explanation and the explanation given by the subjects themselves. But the obvious explanation of this discrepancy, surely, is that people's awareness of their own mental states is imperfect and apt to be influenced, as, notoriously, their perceptual judgements are influenced, by their expectations and preconceptions. And this explanation poses no threat to the folk psychological framework.

And this is the explanation which Nisbett and Wilson themselves propose. The first sentence of the abstract of their paper makes it clear that it is not the legitimacy of beliefs, but the reliability of introspection, with which they are concerned: "Evidence is reviewed which suggests *there may be little or no direct introspective access to higher-order cognitive processes*";[35] and the point is repeated throughout the paper. Wilson's subsequent paper, the one suggesting the "two systems" model, is also clearly concerned with the limits of introspection. What Wilson refers to as the second system, the system which "mediates verbal reports and explanations," turns out to be, specifically, those mechanisms by which people arrive at reports and explanations of their own mental states and processes, and not, as Stich supposes, the system responsible for all verbal behavior whatsoever. Any doubts about the status of Stich's suggestion that Wilson is hypothesising that there are no such things as beliefs are surely finally dispelled by Wilson's own description of the "two systems" model: "A...model will be offered to explain the origin *of beliefs about mental states*";[36] or, indeed, by the subtitle of his paper: "the origins and accuracy *of beliefs about one's own mental states*."

This is just as well. For on Stich's interpretation Wilson's position would be self-defeating. The dis-

crepancy of which an account is required is between subjects' own explanations of the time they took to get to sleep, and the hypothesised true explanation. And the hypothesised true explanation *itself refers to the subjects' beliefs*: arousal subjects take less time to get to sleep *because they believe* that their arousal symptoms are caused by the pills, while relaxation subjects take more time to get to sleep *because they believe* that their thoughts must be especially disturbing.[37]

Stich's other specific argument goes as follows: that if what he calls the "modularity assumption" is false, it follows that there are no such things as beliefs; and that recent developments in AI indicate that the possibility that the modularity assumption is false has to be taken seriously. A system is modular, Stich explains, "*to the extent that there is some more or less isolatable part of the system which plays or would play the central role in a typical causal history leading to the utterance of a sentence.*"[38] Two features of this characterization already raise some doubts about the suggestion that failure of modularity would show that people don't really have beliefs. It makes modularity a matter of degree; whereas, presumably, it is either true or else false that people have beliefs. And it explains modularity in terms of only one kind of behavior, utterances of sentences, ignoring both other kinds of verbal behavior such as assent to sentences, and non-verbal behavior; whereas *a propos* of Nisbett and Wilson Stich himself rightly took beliefs to be supposed to play a role in the explanation of non-verbal as well as verbal behavior. In view of this it is not altogether surprising that in due course Stich's argument seems to disengage from his characterization of modularity. Stich begins, though, by conceding that many cognitive scientists have worked and are working with highly modular models of cognitive processing: he mentions McCarthy's model, which represents memory-storage by a list of sentence-like structures; and Anderson and Bower's model, which postulates a network of nodes, representing concepts, and links, representing relations between concepts, and which he concedes is "still pretty far over to the modular end of the spectrum."[39] But recently, Stich reports, some major figures in the field have begun to favour non-modular approaches: he mentions specifically Winograd, as having formerly favoured modular systems but beginning of late to move away from them; and Minsky, who in his "Frames" paper urges a shift from models which represent knowledge as collections of simple, separate fragments, and subsequently, in his "K-Lines" piece, violates modularity in a dramatic way.

For the present I shall leave Minsky's "K-Lines" paper out of consideration; because in that paper Minsky seems to be proposing a model for *infantile* memory only, and expressly warns that typical *adult* memory processes can be expected to be significantly different. So, though in this paper Minsky does indeed postulate great webs of structure no parts of which correspond to verbally expressible beliefs, it is questionable whether it is of any relevance to Stich's thesis.[40] I shall concentrate on Minsky's "Frames" paper, which is not restricted to infantile memory. Minsky offers only a very imprecise sketch of what frames are supposed to be—am I the only reader whose heart sinks to be told they are like Kuhnian paradigms? or who is hard pressed, in view of Stich's strictures about the lack of clear criteria of identity for beliefs, to refrain from alluding to motes and beams? Frames are described as "data-structures for representing a stereotyped situation, like being in a certain kind of living room, or going to a child's birthday party" and as having "attached" to them several kinds of information, e.g., "about how to use the frame" or "about what one can expect to happen next." Nevertheless, it is clear that Minsky intends his frame models to be *un*like what he calls "logistic" models, i.e., models based on formal, deductive logic, and that one of his objections to logistic models is that they lack a *procedural* element; they specify strings of sentences and permissive rules of inference, but give no guidance as to when which rules are to be used or how the separate bits of information are interconnected. Frames are supposed to be better because they connect bits of information. In a sense, then, it is false that there are isolatable components—in the sense that the components are meant to be *connected.* In another sense, however, it is true that there are isolatable components—in the sense that there are *identifiable parts* of a frame which could plausibly be construed as composed of beliefs; the "child's birthday party" frame, for instance, is plausibly construed as composed of, among other things, the beliefs that guests should bring presents, that party clothes should be worn, and so on.[41] This hardly threatens the legitimacy of beliefs.

Winograd is a trickier case yet. Initially, it seems,

he favoured "declarative" over "procedural" models, but subsequently he came to appreciate the advantages of the latter.[42] "Declarative" here seems to amount to much the same as "logistic" in Minsky's case, and it is already clear that dissatisfaction with this kind of approach need pose no threat to the *bona fides* of belief. However, Winograd seems also to offer an argument of quite another kind, or rather, not so much an argument as a couple of examples introduced by a favourable allusion to a comment of Maturana's, that "many phenomena which for an observer can be described in terms of a representation" can actually be understood as "the activity of a structure-determined system with no mechanism corresponding to a representation."[43] These examples of Winograd's fail to engage with Stich's definition of modularity, for they don't concern "the causal history leading to the utterance of a sentence." But they do have a bearing on Stich's thesis; for they are cases where what looks like goal-directed behavior turns out to be explicable without the postulation of anything like a goal or desire. Stich seems to be best interpreted as inviting one to draw the conclusion that *everything* that looks like goal-directed behavior might be so explicable.

Winograd's first example is a pretty straightforward case in which what might initially seem like goal-directed behavior is plausibly described as really reflexive. There might be some temptation to think of a suckling baby as having a representation of the relevant anatomy, but, he suggests, a better explanation is that it has a reflex to turn its head in response to a touch on the cheek, and a reflex to suck when something touches its mouth—which calls for no ascription of beliefs or desires to the baby.[44] There is no reason to contest Winograd's description of this case; but neither is there any reason to suppose that, of itself, it goes any way towards showing that adults' activities—my going to the fridge to get a glass of milk, for instance, are not sometimes goal-directed. The baby's response—significantly, "behavior" or "action" would be the wrong words—is simple and inflexible in the way characteristic of reflexes, not responsive to circumstances in the way characteristic of what we take to be goal-directed behavior; the suckling baby would react in the same way *whatever* touched its cheek and *whatever* touched its mouth.[45]

But it seems that Stich's point must be precisely that adults' behavior, or rather "behavior," may turn out to be more like babies' responses than we realise. (No doubt this is why he appeals to Minsky's "K-Lines" paper even though it is concerned with infantile memory only. But the appeal is obviously question-begging.) Paul Churchland makes explicitly the claim that is implicit in Stich's argumentative strategy: babies' activities, he says, are recognisably continuous with children's and adults'; so folk psychology, which ascribes beliefs and desires to children and adults but not to infants, makes a distinction where there is no real difference.[46]

An initial reply on behalf of folk psychology or commonsense might be that in one respect, at least, infants *are* unlike children and adults: they can't talk. Since children learn to talk gradually, in a sense there is continuity; but surely our willingness to ascribe beliefs and desires to a small child is apt to run in parallel to his gradual acquisition of language mastery? This is not an accident; for the acquisition of language equips one with resources for representing to oneself connections between possible circumstances, possible actions and possible upshots—a capacity which seems to explain the flexibility and responsiveness to circumstance characteristic of genuine action.[47]

This makes it apparent that Stich is using Winograd's second example to suggest that the flexibility and responsiveness to circumstance characteristic of what we take to be genuine action might after all turn out to be explicable without reference to beliefs or desires. One might be tempted to say of a computer program, Winograd writes, that it has the goal of minimising the number of jobs on the waiting queue; but it is unlikely that it has a goal structure in memory, more likely that "there are dozens or even hundreds of places throughout the code where specific actions are taken, the net effect of which is being described."[48] Presumably the implication is meant to be that, as in the first case it is false that the baby has a representation of the relevant anatomy, so here it is false that the computer has the goal of minimising the job queue.

Winograd is not generous with details of what we are supposed to imagine the computer to do, or what we are to imagine the explanation of its "behavior" to be. Presumably the computer program should be supposed not simply to minimise the queue, but also to manifest something analogous to the other behavior typical of a human agent who, as we suppose, wants to minimise the job queue, such as responding

with irritation to events which impede his efforts to minimise the queue, expressing regret when he misses an opportunity to minimise the queue, explaining when asked what he is trying to do that he wants to minimise the queue, and so on. So we are to imagine a computer which simulates behavior of the same kind of complexity and flexibility: the screen reads "DAMN!" and a robot-arm punches the operator in the nose if he types in a new job just as the computer has processed the last one in the queue; in response to the typed-in command "LIST PROGRAM PRIORITIES" the screen reads, "(1) MINIMISE JOB QUEUE, (2) RECORD JOB TIMES, (3)...," and so forth. And presumably the explanation of the computer's "behavior" should be supposed not only not to require the presence of a simple command like "MINIMISE JOB QUEUE," but also not to require even the presence of more specific commands like "MINIMISE JOB THROUGHPUT TIME," "DELETE IMPROPERLY WORDED JOB SPECIFICATIONS," etc., of which minimising the job queue might be the unspecified upshot. Of course, if, as we supposed, the computer reports minimising the queue as one of its priorities, it must have some representation of this upshot; but it is to be assumed that this plays no role in the explanation of its queue-minimising "behavior."

Suppose such a case. (I have no idea how likely it is, but will not challenge Winograd's judgement on this.) Stich draws the conclusion that adult human behavior might similarly be explicable without reference to agents' beliefs or desires. No doubt this is, in some attenuated sense, *possible*. But how likely is it? The point isn't just that there is a significant logical gap between "this is an effective computer simulation of such-and-such human behavior" and "this is how such-and-such human behavior is brought about"; or that the extrapolation of Winograd's example to a computer which simulates something like the full repertoire of behavior of an adult human being boggles the imagination—though both are worthy of note. It is chiefly that the proposed possible explanation of human behavior—that it is all "structure-determined," that it is not mediated by our capacity for language use—is extraordinarily far-fetched. At any rate, it seems to me that an explanation which couples the versatility and flexibility of adult human behavior with the capacity of language users to represent possible-but-not-actual circumstances to themselves makes better evolutionary sense than an explanation which makes language use no more than an epiphenomenon. (If this makes me a *Cartesian* naturalist, so be it.)

Indeed, Stich's picture seems to make language use, not just an epiphenomenon, but an inexplicable epiphenomenon. For it is something of a mystery what, according to Stich, is going on when, as from time to time they undeniably do, people utter sentences. Which leads me to my final comment. Stich's and the Churchlands' repudiation of beliefs leaves them, as Rorty's repudiation of the objectivity of epistemic evaluation left him, in a curious and uneasy position. What, if it were true that there are no such things as beliefs, could be the status of their utterances (of e.g., the sentence "There are no such things as beliefs")? It is tempting to say that, if there are no such things as beliefs, they don't believe what they say, and so their assertions must be insincere.[49] But it is better to put it this way: if there were no such things as beliefs, our whole notion of assertion, sincere or insincere, would be inapplicable. Proponents of the "no-belief" thesis stand in urgent need of a reconstructed, post-revolutionary concept to replace the old notion of assertion. No such concept, however, seems to be on offer. No wonder, then, that Stich and the Churchlands look so much as if they are kicking away the ladder up which they are climbing.[50]

University of Warwick

Received September 28, 1989

NOTES

1. Traditionalist apriorism is represented throughout the history of epistemology, at least from Descartes through Russell. It has recently been defended in Laurence BonJour, *The Structure of Empirical Knowledge* (Cambridge, Massachusetts: 1985) and George Bealer, "The Boundary Between Philosophy and Cognitive Science," *The Journal of Philosophy*, vol. 84 (1987), pp. 553-55.

2. Jean Piaget, *Genetic Epistemology* (New York: 1970) and Donald T. Campbell, "Neurological Embodiments of Belief

and Gaps in the Fit of Phenomena to Noumena," in Abner Shimony and Deborah Nails, eds., *Naturalistic Epistemology* (Dordrecht: 1987), pp. 165-92, may with a certain amount of rational reconstruction be classified as expansionist naturalists; Alvin Goldman, *Epistemology and Cognition* (Cambridge, Massachusetts, and London: 1986), as scientistic reformist naturalist; Barry Barnes, *Scientific Knowledge and Sociological Theory* (London: 1974) and Barry Barnes and David Bloor, "Relativism, Rationalism and the Sociology of Knowledge," in Martin Hollis and Steven Lukes (eds.), *Rationality and Relativism* (Oxford: 1982), pp. 21-47, as revolutionary naturalists of a sociological stripe; Stephen Stich, *From Folk Psychology to Cognitive Science: The Case Against Belief* (Cambridge, Massachusetts, and London: 1983), Paul Churchland, *Scientific Realism and the Plasticity of Mind* (Cambridge: 1979) and "Eliminative Materialism and Propositional Attitudes," *The Journal of Philosophy*, vol. 78 (1981), pp. 67-89, and Patricia Smith Churchland, "Epistemology in the Age of Neuroscience," *The Journal of Philosophy*, Vol. 84 (1987), pp. 544-53, as revolutionary naturalists of a cognitive-scientific stripe. Manley Thompson, "Epistemic Priority, Analytic Truth and Naturalized Epistemology," *American Philosophical Quarterly*, vol. 18 (1981), pp. 1-12, offers a characterization which corresponds to my category of modest naturalism.

3. Richard Rorty, *Philosophy and the Mirror of Nature* (Princeton: 1979) and *The Consequences of Pragmatism* (Hassocks, Sussex: 1982).

4. E.g., Leonard Nelson, "The Impossibility of a 'Theory of Knowledge,'" first published in German in 1908, reprinted in English in *Socratic Method and Critical Philosophy*, trans. Thomas K. Brown III (New York: 1949), pp. 185-205.

5. Susan Haack, "Rebuilding the Ship While Sailing on the Water," in Robert Barrett and Roger Gibson, eds., *Perspectives on Quine* (Oxford: 1989), pp. 111-27.

6. W. V. Quine, "Epistemology Naturalized," in *Ontological Relativity and Other Essays* (New York: 1969), pp. 68-90; see also his "The Nature of Natural Knowledge," in Samuel Guttenplan, ed., *Mind and Language* (Oxford: 1975), pp. 67-81, and "Five Milestones of Empiricism," in *Theories and Things* (Cambridge, Massachusetts: 1981), pp. 67-72. Hilary Putnam, "Why Reason Can't be Naturalized," *Synthese*, vol. 52 (1982), pp. 3-23, also, I gather, finds Quine ambivalent between reformist and revolutionary attitudes.

7. Rorty's critique of epistemology is discussed at length in David Papineau, "Is Epistemology Dead?," *Proceedings of the Aristotelian Society*, vol. 82 (1982), pp. 129-42; Stich's and Paul Churchland's by Terence Horgan and James Woodward, "Folk Psychology is Here to Stay," *The Philosophical Review*, vol. 94 (1985), pp. 197-226. My approach will be different from theirs, but of course there will be points of contact, mostly, as it turns out, with Horgan and Woodward.

8. Cf. Ernest Sosa, "Nature Unmirrored, Epistemology Naturalized," *Synthese*, vol. 55 (1983), pp. 49-72, especially p. 70.

9. On the history of epistemology, see Rorty, *Philosophy and the Mirror of Nature*, *op. cit.*, pp. 131-36; on the optical metaphor, *ibid.*, pp. 146, 159, 162-63.

10. Rorty, *Philosophy and the Mirror of Nature*, *op. cit.*, pp. 155-64, 168-212.

11. For the conversational theory, see Rorty *Philosophy and the Mirror of Nature*, *op. cit.*, pp. 170, 175-76, 315ff.; for incommensurability, *ibid.*, pp. 315-33; for the critique of correspondence, *ibid.*, pp. 284-311, 333-42, 377, and *The Consequences of Pragmatism*, *op. cit.*, introduction.

12. *Philosophy and the Mirror of Nature*, *op. cit.*, pp. 315-22; *The Consequences of Pragmatism*, *op. cit.*, introduction, especially section 5.

13. But I note that there do seem to be continuities between modern epistemology and some problems that go back at least as far as Descartes; and that it is questionable whether Rorty's account of the role of the optical metaphor is compatible with the non-cognitivist theory of metaphor he elsewhere defends. See his "Unfamiliar Noises: Hesse and Davidson on Metaphor," *Proceedings of the Aristotelian Society*, Supplement, vol. 61 (1987), pp. 283-96; and cf. Susan Haack, "Surprising Noises: Rorty and Hesse on Metaphor," *Proceedings of the Aristotelian Society*, vol. 87 (1987-88), pp. 179-87.

14. David Armstrong, *Belief, Truth and Knowledge* (Cambridge: 1973), is *perhaps* an example, though it is not altogether clear that it is correct to construe him as offering a theory of justification at all.

15. Cf. my "Rebuilding the Ship," *op. cit.*, section I.

16. Wilfred Sellars, "Empiricism and the Philosophy of Mind," in *Science, Perception and Reality* (London: 1963), pp. 127-96.

17. Donald Davidson, "A Coherence Theory of Truth and Knowledge," in Dieter Henrich, ed., *Kant oder Hegel?* (Stuttgart: 1983), pp. 423-38; I note that Davidson, using a very broad characterization, classifies Rorty as a coherentist. See also Karl Popper, *The Logic of Scientific Discovery* (London: 1959), chapter 5, for an early statement of the "confusion of

justification with causation" argument, there used in a critique of "psychologism."

18. The term was introduced and the theory sketched in my "Theories of Knowledge: an Analytic Framework," *Proceedings of the Aristotelian Society*, vol. 82 (1982-83), pp. 143-57; see also "Rebuilding the Ship," *op. cit.*, section I.

19. W. V. Quine, "Two Dogmas of Empiricism," in *From a Logical Point of View* (New York and Evanston: 1953), pp. 20-46 in revised, 1961, edition.

20. Cf. "Rebuilding the Ship," *op. cit.*, section II.

21. See my "Realism," *Synthese*, vol. 73 (1987), pp. 275-99 for amplification of the distinctions made in this and the previous paragraphs. It should be obvious that I would dispute Rorty's classification of his view as "pragmatist."

22. Since Rorty downplays it, I have not considered the role played in Kuhn's arguments for incommensurability by the idea of meaning-variance. But cf. "'Realism,'" *op. cit.*, section III, where I distinguish stronger and weaker forms of "the" meaning-variance thesis, and argue that the usual arguments support only the weaker forms, but only the strongest has radical epistemological consequences.

23. For the first suggestion, see *Philosophy and the Mirror of Nature*, *op. cit.*, p. 343, and *The Consequences of Pragmatism*, *op. cit.*, p. xxxvii; for the second, *Mirror*, p. 320; for the third, *ibid.*, pp. 377-78.

24. Stich, *From Folk Psychology to Cognitive Science*, *op. cit.*, p. 2. Stich is aware that epistemology is not the only discipline threatened if there are no beliefs; he mentions history and economics as other potential casualties. I note that the whole institution of the law would also come under threat.

25. Churchland, "Epistemology in the Age of Neuroscience," *op. cit.*, pp. 544-45, 546.

26. Churchland, "Epistemology in the Age of Neuroscience," *op. cit.*, p. 547.

27. Karl Popper, "Epistemology Without a Knowing Subject," *Objective Knowledge* (Oxford, 1972), pp. 106-52; Susan Haack, "Epistemology *With* a Knowing Subject," *Review of Metaphysics*, vol. 33 (1979), pp. 309-35; "Rebuilding the Ship," *op. cit.*, section I; "What is 'the Problem of the Empirical Basis,' and Does Johnny Wideawake solve it?," delivered to the British Society for the Philosophy of Science, 1988 forthcoming, *British Journal for the Philosophy of Science.*

28. See Quine, "Two Dogmas," *op. cit.*, section 6, and cf. "Mind and Verbal Dispositions," in Guttenplan, ed., *Mind and Language*, *op. cit.*, pp. 83-95.

29. Stich, *From Folk Psychology to Cognitive Science*, *op. cit.*, pp. 229-30.

30. Cf. Horgan and Woodward, "Folk Psychology is Here to Stay," *op. cit.*, pp. 198-99. Stich is aware of the possibility that sentences of the form "A believes that *p*" may be admitted as true without postulating an ontology of beliefs. But since my account of justification requires the postulation of *S*-beliefs (belief states) and C-beliefs (belief contents), this is not a possibility on which I am disposed to dwell.

31. Churchland, *Scientific Realism and the Plasticity of Mind*, *op. cit.*, sections 12-16; "Eliminative Materialism and Propositional Attitudes," *op. cit.*, section II.

32. Churchland, *Scientific Realism and the Plasticity of Mind*, *op. cit.*, pp. 39-41, 96-100.

33. Cf. Horgan and Woodward, "Folk Psychology is Here to Stay," *op. cit.*, pp. 203-10.

34. Stich, *From Folk Psychology to Cognitive Science*, *op. cit.*, pp. 230-37; Richard Nisbett and Timothy Wilson, "Telling More Than We Can Know: Verbal Reports on Mental Processes," *The Psychological Review*, vol. 84.3 (1977), pp. 321-59; Timothy Wilson, "Strangers to Ourselves: the Origins and Accuracy of Beliefs About One's Own Mental States," in J. H. Harvey and G. Weary, eds., *Attribution: Basic Issues and Applications* (Orlando: 1985), pp. 1-35.

35. Nisbett and Wilson, "Telling More Than We Can Know," *op. cit.*, p. 231.

36. Wilson, "Strangers to Ourselves," *op. cit.*, p. 16, my italics.

37. Cf. Horgan and Woodward, "Folk Psychology is Here to Stay," *op. cit.*, p. 208.

38. Stich, *From Folk Psychology to Cognitive Science*, *op. cit.*, p. 238.

39. Stich, *From Folk Psychology to Cognitive Science*, *op. cit.*, pp. 238-39.

40. Marvin Minsky, "K-Lines: a Theory of Memory," in D. Norman, ed., *Perspectives on Cognitive Science* (Norwood: 1981), pp. 87-103; the quotation is from p. 100. Of this paper, Minsky writes, "The idea proposed here—of a primitive 'disposition representing' structure—would probably serve only for a rather infantile dispositional memory; the theory

does not go very far toward supporting the more familiar kinds of cognitive constructs we know as adults...I doubt that human memory has the same, uniform, invariant character throughout development" (p. 88).

41. Marvin Minsky, "Frame-System Theory," in P. Wason and P. Johnson-Laird, eds., *Thinking* (Cambridge: 1977), pp. 375-76; "A Framework for Representing Knowledge," in John Haugeland (ed.), *Mind Design* (Cambridge, Massachusetts: 1981), pp. 94-128. For the characterization of frames and the allusion to Kuhn, see "A Framework for Representing Knowledge," pp. 96-97; for the contrast with logistic models, *ibid.*, pp. 123-28.

42. Terence Winograd, "Frame Representations and the Declarative-Procedural Controversy," in D. G. Bobrow and A. Collins, eds., *Representation and Understanding* (San Francisco, New York and London: 1975), pp. 188-210.

43. Terence Winograd, "What Does it Mean to Understand Language?," in Norman, *Perspectives on Cognitive Science*, *op. cit.*, pp. 231-63; the discussion of Maturana is on pp. 248ff. Unfortunately Winograd's reference for the quotation from Maturana seems to be mistaken, and I have not been able to locate the source.

44. Winograd, "What Does It Mean to Understand Language?" *op. cit.*, p. 249.

45. Cf. D. E. Woolridge, *The Machinery of the Brain* (New York: 1963), cited in Daniel Dennett, *Brainstorms* (Hassocks, Sussex: 1979), pp. 65-66, on the "behavior" of the *sphex* wasp.

46. Churchland, *Scientific Realism and the Plasticity of Mind*, *op. cit.*, p. 249.

47. Commonsense psychology does not of course deny that some things adults do may be explicable by reference to inbuilt responses to certain stimuli; or that we may sometimes be mistaken about the explanation of our own behavior; or that some of the things we do are the unintended, sometimes the undesired, consequences of things we desire to do. Note also that I do not claim that only language-using creatures can engage in goal-directed activity; think of chimpanzees reaching for inaccessible bananas with a stick, for instance. But surely possession of a language greatly increases our capacity for goal-directed behavior.

48. Winograd, "What Does it Mean to Understand Language?," *op. cit.*, p. 250.

49. Patricia Churchland discusses this version of the objection in "Is Determinism Self-Refuting?," *Mind*, vol. XC (1981), p. 100.

50. This paper was read, in various versions, at the Ockham Society, Oxford, Howard Robinson replying; the University of California, Riverside; Florida State University, Tallahassee; the University of Miami; and University College, London. I am grateful for comments and questions on these occasions; and to John Barker, Stewart Granger and Mark Migotti for helpful correspondence.

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 27, Number 3, July 1990

THE ANATOMY OF AGGRESSION

Steven Luper-Foy

...Men are continually in competition for honour and dignity...; and consequently amongst men there ariseth on that ground, envy and hatred, and finally war....

Thomas Hobbes, *Leviathan*

QUITE mundane pursuits as well as lofty attempts to achieve the extraordinary turn us against each other in tragic, insidious ways. These pursuits give rise to an "invisible hand" that, far from guiding people toward happiness, steers them instead toward confrontation and aggression. People end up literally making war in order to secure a good life. My aim here is to lay bare mechanisms by which our undertakings make aggressors of us. I begin with an analysis of competition, aggression, and related phenomena.

COMPETITION

Competition pits individual against individual or group against group. Games involving winners and losers, such as poker and races, are relatively low-key cases of competition; others, such as boxing matches and wars for territory, are more dangerous. The common element in these cases is rivalry, the struggle among agents trying to outdo each other in some pursuit. In fact, people may be said to be *competing* with each other just in case there is some item *X* that each seeks, and each, in pursuing *X*, is aware of taking (or being prepared to take) steps which make it more difficult for the others to attain *X*. By contrast, people are *cooperating* just in case there is some item *Y* that each seeks, and each, in pursuing *Y*, is aware of taking (or being prepared to take) steps which *help* the others to attain *Y*.

When rivals confront each other, they are after some *competitive property*, a property with the following feature: where along a dimension an item is required to fall in order to have that property depends on where along it other items of the same sort actually fall. What it takes for a person to qualify as *best surgeon*, for instance, depends on where other surgeons fall along the dimension of surgical expertise. Two further examples of competitive properties are *being unique, or exceptional, in some given respect*. These are also properties that not everyone could *possibly* have; they are *necessarily nonuniversalizable*. Other competitive properties are only *contingently nonuniversalizable* they are nonuniversalizable only as a matter of fact, and could be possessed by everyone were the world different. An example: *exclusively owning a home in Manhattan*.

The members of a group of people will compete with each other just in case each is sufficiently motivated to secure the same nonuniversalizable competitive property. If a group of us all aim at *being the best writer of the group*, a necessarily nonuniversalizable competitive property, then any progress one of us makes toward that goal will constitute a setback for at least one of the others, and the success of anyone will require the defeat of everyone else. Similarly, desiring the (contingently nonuniversalizable) property of *being well-fed* will make competitors of us when food is scarce. In seeking a nonuniversalizable competitive property, people become rivals, and hence competitors. But some competitive properties need not make rivals of their pursuers. Many such properties are actually universalizable; an example is *being unique in at least some respect*. People who acquire universalizable properties may do so without interfering with the attempts of others to reach the same goal, and hence competition need not result.

Anytime a group is competing, the desire for a *nonuniversalizable* competitive property is motivating its members. Still, a situation in which we are

each pursuing (competing for) the *one* goal of being well fed can just as well be described as one in which *several* goals are being pursued: *my* desire is that *I* be well fed, while yours is that *you* be well fed, and similarly with the other people. Thus there is a sense in which we each have our separate desires rather than a common goal. Yet each of our desires is an indexical variant of the others: each expresses the goal that an individual end up with a specific competitive property, in this case *being well fed.*

Already we can see that most of us have aspirations that bring us into competition with other people. In turn, this competition frequently generates aggressive and evil behavior. How it does so I shall discuss after making a few points about the nature of evil and aggression.

Evil and Aggression

The most seriously immoral acts are those of people who aim to cause misfortune and do so because they believe it to be intrinsically valuable. What such people regard as intrinsically good is *misfortune*, which, I shall assume, involves substantial harm; the acts by which they pursue their goal are therefore irremediably evil.

Acts that are less serious may be at least prima facie evil as well. The point of the qualifier "prima facie" is that in unusual circumstances a prima facie evil act may be morally permissible. It is possible for us to do something designed to harm others substantially even though we do not regard the worsening of their prospects as good in itself. To *cause* others a misfortune is, of course, just to cause a situation which itself constitutes a misfortune for them. By contrast, our act is *designed* to cause them a particular misfortune, *M*, if and only if we want to cause *M because* we think *M* is a misfortune-constituting situation. Hence planning to cause someone a particular misfortune entails planning to bring about a misfortune-constituting situation, but it is possible that we value producing that situation entirely for instrumental reasons rather than because we think it intrinsically valuable. If an adolescent destroys someone's prized car solely because his peers have pressured him into finding some way to create misery for another, and he wants to avoid becoming the object of ridicule, then he has designedly harmed another without regarding it as intrinsically valuable. Nonetheless, his act is prima facie evil, and almost certainly unqualifiedly evil.

All acts that are designed to cause misfortune are serious enough to be prima facie evil. The same is true of acts that cause anticipated but unintended and unwanted misfortune. If I decide that the only way to save my starving children is to rob you of your life savings, I know that I am causing you a grave loss, but the fact that I anticipate your loss does not mean I intend or want it. I would prefer that you not undergo a misfortune, and would have robbed you even if it were not a loss for you. While such acts (that cause anticipated but unintended harm) are, ceteris paribus, immoral, nevertheless acts intended to cause harm are more serious than ones that cause unintended but foreseen misfortune, which in turn are worse than ones causing anticipat*able* (but unforeseen) misfortune.

These prima facie evil acts, including the last and least serious, are all acts of aggression. For "aggression" refers to acts intended to cause significant harm to others, and even to acts that cause anticipatible misfortune. When the harm we cause is reasonably easy to predict we are still aggressors even if we did not make the prediction, say because we wanted to accomplish some goal so badly that we ignored the suffering we would cause others. A great deal of self-deception is involved in the way people represent to themselves the contexts in which they pursue their ends. There are a thousand ways to rationalize even those acts whose harmful consequences we admit to ourselves; and where rationalization fails, we can refuse to admit to ourselves the unsavoriness of the means we use to pursue our ends. I see no reason to withhold the epithet "aggressor" from those who delude themselves about the pain for which they are responsible. Anticipatible misfortune ought to be anticipated; those who neglect to think out the consequences of their acts are still culpable.

Thus on my account acts of aggression may or may not be evil, and while there is always a prima facie case against them, that case sometimes will be outweighed. When aggression is the only means to prevent much greater evil, when it is required to rectify an injustice, and when the victim consents to (risk) the misfortune, aggression is often permissible.

Only acts, not omissions, can constitute aggression. A physician's decision not to give a new life-saving drug to patients who subsequently die does

not constitute aggression even if the drug is easily available. Not administering a drug is an omission. By contrast, blocking access to a life-saving drug is an act, and constitutes aggression.

My use of the term "aggression" is in tension with that given it by theorists who maintain that aggression is innate. According to innatists, animals like fishes are capable of aggression, and observations about the behavior of animals constitute the main evidence for the nativist thesis. Few if any animals other than people know what constitutes a misfortune, and so few could anticipate harm, though many kinds of animals have an inborn tendency to attack members of their own species in certain circumstances. Of all animals, then, human beings seem uniquely capable of aggression in my sense of the term.

Whether and in what sense a tendency, instinct or drive to attack others is an innate feature of human beings is an important, politically-charged issue. The literature on the topic is extensive but in my opinion far from conclusive.[1] Culture inspires so many motivations which themselves would lead to attack behavior that it seems premature to posit an innate tendency to attack others. Thus greed, ambition, and "competition for honour and dignity" (to use Hobbes' phrase) could easily inspire attacks on others; I have seen no one argue that they are innate drives, however, and it is far-fetched to say that they are themselves inspired by a (sublimated?) urge to attack others since they are better suited to explain that urge than the latter is to explain *them*. Fortunately, I need not resolve the innatism issue because my focus is on describing certain sorts of goal that generate the desire to cause misfortune, and these desires could be innate, acquired, or themselves the product of innate drives. So my account is compatible with both nativism and antinativism.

The characterization of aggression as acts causing anticipatible misfortune is suggestive. If correct, then a key to aggression is understanding why people come to value the misfortune of others. I will suggest that they do so for the same reasons that they value successful competition.

The Misfortune of Others

Suppose that you and I and several others attribute intrinsic value or at least great importance to being the fastest runner of the group. Then we may or may not undertake a *contest* to determine who is fastest; indeed, if being fastest is very important to us, yet we are unsure of our skill, we may prefer to avoid a contest, thereby avoiding the risk of losing. But we will still compete. For we will seek the means to beat the others, by endless jogging, eating certain foods, etc. Assuming that we are aware of the value the others attribute to being fastest, and that we do not deceive ourselves, then we will realize that any significant steps we take in our competition constitute misfortune for the others, for substantial progress we make toward being fastest makes it significantly harder for the others to achieve something important to them, which is a misfortune for them. Hence we are aggressors.

The *mere* anticipatability of the fact that we are about to harm others would make us the least objectionable sort of aggressor. More serious aggression is generated by competitions whose participants *want* to cause each other misfortune. Precisely this occurs when our interest in winning comes to be linked to a concern that *their* winning be important to others. This linkage will occur when people who consider being fastest intrinsically valuable (or at least very important) decide to race as a means of obtaining the competitive property of *being proven the speediest*. If you and I form such a group, then the steps I take not only cause you misfortune, but are designed to do so, and hence (unless mitigating circumstances obtain) constitute evil—similarly for you and the others. I know that you want to win, and, more importantly, I *want* you to want to win, and similarly for you. The contest is useless to me as an indicator of who is fastest unless everyone tries to win, and that ordinarily requires that they want to win. Moreover, I want it to be the case that your losing would be a misfortune for you: since I desire that you want very badly to win, then I shall desire that your losing will constitute an important setback for you, which is to say that I want it to be a misfortune for you. And of course your losing *would* be a misfortune for you. You consider winning important since you attribute intrinsic value to being fastest, and failing to achieve something of importance to you would constitute a misfortune for you. Finally, in wanting to win I also want specifically to *defeat* your desire to win, which, we said, I want to constitute a misfortune for you. Unless extenuating factors exist, we can conclude that the steps I take toward

defeating you constitute evil. So do your steps toward my defeat.

Suppose now that we attribute intrinsic value specifically to winning a contest that is important to its participants. Winning is inseparable from defeating others, so in attributing intrinsic value to winning contests whose outcomes matter greatly to their participants, I attribute it also to defeating aims the others consider important. In such cases it is all the more obvious that my aggressive competition is prima facie evil.

But some competitions generate neither evil nor aggression. Consider ones whose participants consider winning a relatively trivial matter. A lighthearted hand of cards is an example, as are many other games, particularly ones that are not zero-sum. Winning is not important in such contexts, so losing harms no one, and wishing to defeat an opponent's desire to win is at worst a minor form of maliciousness. By the same token, such competition will not be very heated. Other games have high stakes and hence could be the occasion for aggressive competition, and even evil, as when a demonic madman forces me to play for my life.

I have said that all contests generated by the attribution of substantial importance to nonuniversalizable competitive properties will involve aggression and that some will involve evil. However, these claims do not entail that people who strongly aspire to nonuniversalizable competitive properties *must* end up aggressors. We may not act on *any* particular desire, except perhaps one that is overwhelmingly strong. But of course we *will* strive to get what we badly want unless we have a competing set of goals. Thus we may back out of an opportunity to become top pianist in the area because doing so requires the defeat of a friend who covets the title. Our concern for our friend is strong enough to curb our ambition. In general, to the extent that we have an interest in (avoiding anything detrimental to) the well-being of people in general—or, what is more likely, a certain set of people in particular—we will tend to avoid aggressing against them. Obviously a concern for morality is a complicated version of precisely this interest. Hence the moral presumption against aggression.

Even if we take a strong interest in not harming others, our nonuniversalizable competitive goals are likely to make aggressors of us because our desire to attain nonuniversalizable competitive goals are not always completely outweighed by our desire to avoid harming others. Almost all of us balance between our interest in the well-being of others on the one hand and the pursuit of our own ends on the other even if we ignore or deceive ourselves about what we are doing. We would quite rightly insist on our fair share of scarce natural resources even if it were a misfortune to certain other people (say because they cannot plan children since they cannot expect to feed them) that they do not get our share as well as theirs. And what is in the end our fair share is always elusive, partly because our interests *vis-á-vis* those of others look more important to us than they do to others. *A fortiori*, most people who find themselves in circumstances of extreme scarcity would fight to obtain at least as much food as they need to keep themselves and their families alive.

Even those who give complete priority to avoiding harm to others might well end up aggressors. Saints who would forego any benefit to themselves that would lessen the prospects of another (making saints of people with an overwhelmingly powerful concern for the welfare of others and an underwhelmingly low estimation of their own private ends) might find themselves morally obligated to aggress, as in the situation in which a life-saving drug is developed that is so scarce that access to it by some must be blocked.

Aggression and Worthwhile Lives

Given that the pursuit of nonuniversalizable competitive properties so easily produces aggression, it is unfortunate that precisely this pursuit plays a central role in the lives of many individuals. Wanting to be an entertaining novelist, profound philosopher, enchanting artist or informative historian or scientist is one thing; wishing to be an exceptional or unique novelist, artist, philosopher, historian, or scientists is quite another. That the desire for such nonuniversalizable properties is widespread was noted by Thomas Hobbes in *Leviathan* (I, 8), where he emphasized the prevalence of the passion for honor and for power of all sorts. Alfred Adler went so far as to say that "the striving to be superior is innate."[2] Nietzsche's emphasis on competitive properties is well known. And Abraham Maslow is famous for claiming that after our basic needs are met various "higher" desires become felt; at least many of these turn out to be desires for competitive properties,

according to Maslow's picture. Adler and Maslow thought of these as innate desires that are the root of our urge to ennoble ourselves, so that achieving "higher" goals is an innate desire of us all. While the nativist thesis is undoubtedly an exaggeration, it is clear that desires for competitive properties are considered critically important by many people. Hence it is important to notice that these seemingly noble pursuits create misery. Let us explore this fact in more detail.

More often than not, people do whatever is necessary to save their lives, however loath they are to do so, since they are still more loath to die. But just as we expect people to defend themselves, so we should expect them to go to great lengths to protect things whose importance is about as great as the value people assign their lives. Indeed, many of us will face certain death when the lives of our offspring, spouses, or close friends are in danger. The continued welfare of these loved ones is often part of what people consider to be the minimal requirements of a worthwhile life.

The minimal requirements of a worthwhile life (for a given person) are the conditions which (according to that person) must be satisfied in order for life to be at least minimally worth living, so that if the requirements are not satisfied, then living is considered no better than being dead. Desires whose satisfaction we consider requisite to a minimally worthwhile life are capable of leading us to put our lives in the most serious peril. To come between us and our attempt to satisfy them is tantamount to threatening our lives: if we find ourselves with the belief that our lives are and will remain unworthwhile (and the view that this belief will remain unshakeable), we have no reason to persist in life. Hence, the closer a goal comes to being part of what we consider to be a minimally worthwhile life, the more extreme the means we will be willing to use to achieve it. For this reason, the search by some people for a meaningful life takes on a special desperation. Such people are trying to identify what the components of a worthwhile life *would be*, something they believe they must do before they can achieve a worthwhile life.

Now notice what happens when people consider aggression-generating pursuits to be central to a worthwhile life. The pursuit of nonuniversal competitive goals involves people in activities which undermine the attempts of others to fulfill similar goals. Hence each of us is led to view the attempts of others to flourish—to fashion worthwhile lives for themselves—as an aggressive attack on us. And the importance we assign to flourishing leads us to return the aggression, even if we notice that doing so makes it more difficult or impossible for others to flourish. As long as we continue to assign central importance to nonuniversalizable competitive goals, the bitter struggle among us will remain.

War and Collective Aggression

The worst forms of violence are produced by clashes among *groups*, not individuals. When such clashes occur, we speak of war; in its primary application the term refers to violent clashes among entire groups, entire collectivities, not confrontations, however violent, among individuals. It is sometimes even said[3] that wars must involve entire nations, but I see no reason to focus on this unit. Since war certainly antedates (and is probably in some measure responsible for[4]) the emergence of the nation-state, the definition is unduly restrictive.

I cannot attempt to describe all of the reasons people go to war. However, I do want to emphasize that aggression generated by the pursuit of competitive properties plays an important role in the emergence of warfare. Just as individuals compete against others, so groups compete against others. Individuals derive a sense of worth not just from their own accomplishments but also from the accomplishments of groups with which they identify, and they can lose their sense of worth from the failures of these groups. The great importance we place on our collective accomplishments is revealed by such phenomena as the fact that a team member who played extremely well can still be crushed at the failure of the team as a whole. It is the high estimation of collective worth as opposed to individual worth that is lost by such an individual, the judged worth of *us* rather than the judged worth of *me*. But the collective worth we feel as members of groups with which we identify is as important to many of us as individual worth.

Group conflicts are especially dangerous, and not only for the simple reason that many people lend a hand in the violence. Another reason is that as a rule people who identify with a collectivity consider themselves and are considered by others as less important than the whole. Rough criteria of group

identity enable people to say that the group survives the deaths of individuals, and so the latter tend to become expendable components of the whole. This device was exploited by Hitler, who, by requiring that party members swear loyalty to *him*, ensured that he was the only group member who *wasn't* expendable.

Getting us to subordinate our own interests to those of the nation (by encouraging us to identify with the national group) is one of the thrusts of nationalism, or national patriotism. Another is getting us to subordinate the interests of other national groups to those of our own national group. The upshot is that even people who assign a great deal of weight to pursuing their own interests only in ways that do not harm others are encouraged to think that harm to those outside the nation is comparatively trivial. Limiting our concern for others to those in our group obviously makes war all the more likely by making us more unwilling to compromise.

Defeating the important aspirations of groups can therefore constitute a misfortune for the individuals who not only consider themselves part of those groups but who also judge the group's survival according to various criteria of collective identity. Hence aggression can occur on the group level, and *will* to the extent that groups value and pursue nonuniversalizable competitive properties.

Positional Goods

There is a type of goal that generates a different form of aggression than that discussed so far. Consider a goal which people *want* to achieve only if others fail to achieve it (or an appropriate indexical variant), or (to characterize the goal more usefully) a desire for an item which people value possessing only to the extent that others do not possess that item. Dress styles are clear examples, at least insofar as they are intended to make us look unusual: this purpose is defeated if the style becomes too popular with others. Prizes are another example. Such things economist Fred Hirsch has called *positional goods*.[5] He also used this term to refer to necessarily nonuniversalizable competitive properties such as status, prestige and greatness. However, the latter properties cannot be described as things we value only to the extent that others lack them. It makes no sense to ask whether we would value them if everyone possessed them since we *can* possess them only if others do not. (Compare the competitive property *equality*: it may be possessed only *if* everyone does.) Nonuniversalizable competitive properties are quite distinct from things we value only to the extent that others lack them. The distinction is so apparent that I shall use Hirsch's term *positional goods* to refer only to the latter, that is, to goods we value more to the extent that less people acquire them.[6] And for obvious reasons the term may be used to cover items we value *less* to the extent that more people acquire them.

Unlike nonuniversalizable competitive desires, then, positional goals *can* be achieved by everyone (in a group). But they are valued only if relatively few achieve them. Still, I suggest, these goals owe their positionality to more fundamental competitive desires, universalizable or not. They follow the model of styles of clothes aimed at making wearers unusual dressers; a particular style loses value when others adopt it because by adopting it people frustrate each other's (nonuniversalizable) goal of being unusual dressers. As this example shows, one reason we might desire or value something X more when we find that others have not attained it (or less when we find that they have) is that our interest in X is based on more fundamental competitive desires. Perhaps we want to find something or other (we are not particular) we can have more of than most or all other people; perhaps we want to come as close as possible to the *exclusive possession* of something or other. When X is scarce, like black coral, any of these underlying (universalizable) desires will lead us to develop an interest specifically in it; we will want to appropriate black coral when we find that it is attained by few others. On the other hand, we will lose interest in black coral to the extent that it is widespread, but we lose interest because it does not allow us to fulfill our underlying competitive desire. Everyone may be free to acquire black coral, but doing so prevents others from using it to fulfill their competitive goals.

If we *already* possess something, say the love of Hilary, or at least think we do, then the interest in the exclusive possession of Hilary's love is one of *jealousy*. On the other hand, if (we think) it is possessed by someone else, not us, then our attitude is one of *envy*. What converts an interest in X into a jealous interest in X is that besides our desire to retain X we have the desire that no one else have it. Once jealousy is added to our interest in Hilary's love, we

have a compound desire concerning it, so that losing it is no longer the only way our desire can be thwarted. Someone else's securing Hilary's love would prevent us from satisfying our compound desire. So if you no longer receive Hilary's love, your jealousy will lead you to want no one else to receive it either.

Envy is susceptible to a parallel analysis. An interest in obtaining someone else's possession *X* is converted into an envious interest when we add our desire that no one except us possess *X*. The resulting compound desire will motivate us to seek *X* out, and also to endeavor that others lose it.[7]

There is, then, little doubt that many positional desires stem from desires to approximate the exclusive possession of something; they are generated by the same interest in exclusivity as gives rise to jealousy. Indeed, I suggest that all positional goals are the result of underlying competitive desires. Moreover, a group's pursuit of the conjunctive desires that are constitutive of envy and jealousy results in straightforward aggression when the stakes are high enough: "our" pursuit of the more or less exclusive possession of *X* directly interferes with "theirs," so if the exclusive possession of *X* is of great importance to them, our acquiring it will constitute a misfortune for them.

But our attempt to fulfill positional desires will generate its *own* form of aggression as well. Earlier when we discussed aggression we spoke of acts which left victims with desires they considered very important but which they could fulfill only with enormous difficulty if at all; it was in this sort of interference that aggression was said to consist. But there is another type of interference, one that gives rise to a second form of aggression. A second sense in which someone can interfere with our attempts to satisfy a desire is to *eliminate* it, to bring about a situation in which our desire, while never satisfied, no longer exists. The most drastic way to do away with one of our desires is naturally to do away with *us*. Less drastically, I could seriously interfere with your attempt to fulfill the goals of your life plan if I brainwash or drug you so as to take away your desire to fulfill those goals. Removing desires whose satisfaction we consider profoundly important to us can constitute aggression since such losses can be misfortunes for us.

This "elimination" type of interference occurs whenever a group of people tries to acquire items which for them are positional goods. For example, as more and more members of a group acquire prestige items like Mercedes, they find that they value them less and less. Even a Mercedes cannot remain a prestige item if everyone has one. The satisfaction of positional desires by many people yields no one any satisfaction. People find that they have devoted time and energy to the acquisition of items whose value is drained away to nothing by the similar efforts of others. Of course, most people do not want to spend time in pursuits which will shortly lose all value, and it is probably typical for people not only to want to fulfill their plans but also to want not to lose the desire to fulfill their plans. To undermine the value their pursuits had, or to eliminate their motivation to achieve their plans, is thus to leave them with an unfulfilled desire. Hence a group's pursuit of positional desires can generate not only the "elimination" form of aggression, but also the "frustration" sort, involving victims with unfulfillable desires.

It might appear that we could arrange things so that the interest in positional goals need not by itself generate aggression. If people were *very* dissimilar in their ambitions, so that they always chose to work toward goals and seek out goods *other than* the ones sought by others, then competition and aggression would not result. People's pursuits would not undermine the value of the things sought by others. A similar point may seem to hold for nonuniversalizable competitive desires: if these varied from person to person in such a way that we could coordinate our pursuits so as to avoid the territory of others, then aggression would not arise. However, these solutions are unavailable, for more than one reason.

First, competitive and positionally-minded people take an interest in how the properties they seek compare to those sought by others. Some of the properties people covet are more prestigious than others, as ranked by competitive criteria such as *possessed by only the most intelligent people*. Hence they would likely react to a situation in which everyone is best in some respect by saying that being best in some respect is relatively valueless since everyone has achieved it; what is truly important is being best with respect to a property that scores high *vis-á-vis*...(fill in the blank with a competitive criterion applicable to properties). (I ignore here the fact that an insincere spirit of egalitarianism sometimes motivates competivists to (pretend to) forego evaluating competitive properties themselves; thus we hear

"It does not matter what you do so long as you do it better than anyone else," and we cannot help but wonder, "If what you do does not matter, why should it matter that you do it better than anyone else?")

Second, competitive people tend to want to have *more* (prestigious) competitive properties than most others possess. Only a few may have more of them than most other people, hence frustration aggression results. Our concern to surpass others in our total package of competitive properties also keeps us interested in securing what they have, and in preventing others from adding items to their package which do not show up in *our* package, so that even commonly distributed possessions retain some value to us. It inspires us to "keep up with the Jones," to *emulate* others, to crave what others crave, and when others accomplish some goal, to become interested in accomplishing our indexical variant of that goal. Yet this mechanism interferes with people's attempts to acquire valued positional goods by ensuring that too many people seek and thus come to possess them. Soon, as more and more people obtain the item, it loses its vulnerable positional value, forcing people to pursue some new sort of item. And so it goes; people constantly struggle to possess or achieve things that few have, and are hotly pursued by would-be emulators.

A consequence of this dynamic is that the things we value which can easily be possessed by many (commodities such as styles of clothes, for example) will tend to change over time as more people attain those things and thus undermine their value. Another consequence is that we tend to adopt ever more unreachable goals. The more unattainable the goal, the more unlikely it is that others will undermine the value of our attaining it by themselves attaining it. As well, easily attainable goals already will have been secured. Of course, certain goals can only be attained by few; this exclusiveness helps ensure that they always will be valued, that quite independently of their intrinsic features they will always hold the greatest prestige. The interest people have in things that are difficult to attain helps explain why people take up such inane pursuits as being the first person to walk across Death Valley (carrying one's own water, or while dragging boulders, or whatever).

A final point: we tend to have positional and emulative desires only against the backdrop of groups of people who take an interest (whether favorable or not) in each other's affairs: in order to emulate, there must be some group we want to emulate; to excel, there must be a group we want to exceed, and in both cases we tend to find it impossible to sustain our interest in our endeavor if the group is indifferent about it.[8] It is worth emphasizing that group members in the grip of competitive desires are *contending against* each other, rather than cooperatively pursuing a common good. Deeply ironic is the fact that the value of excelling depends on there *being* a group of people who show by emulation that they do not want to be surpassed, for in that sense the people who are competing need each other for there to be anything worth striving for, yet *what* they are striving for defeats the aspirations of the people they need. They flourish at the expense of the people who make their flourishing possible.

Against Competitivism

Two powerful sources of aggression and evil, we have seen, are the attribution of importance to and consequent pursuit of nonuniversalizable competitive properties and positional goods. Such pursuits clearly must bear some of the onus of moral impermissibility that is borne by aggression. However, it is by no means clear how this burden gets transferred, nor, indeed, when it is that aggression is wrong. And even when the issues of moral permissibility are sorted out, there remains that of whether it is a *good idea* for people to vie against each other for superiority. In what follows I shall discuss these issues and sketch thereby some of the disadvantages of letting positional and nonuniversalizable competitive values play an important part in our lives.

Let me begin, however, by dealing with an objection: anything that might be said about the disadvantages of nonuniversalizable competitive values is bound to be misleading (according to the objection) since the value of competition itself is so great. Consider for example the benefits people reap from races among scientists each of whom wishes to be the sole discoverer of the first cure for cancer (assuming someone wins). Truly, we would all benefit from a cure for cancer; competition genuinely has some beneficial results. But is competition among researchers really the most effective way to produce those beneficial results? If so, presumably it would be because competition is a more effective motivator than the alternatives: *individual* effort that is not

aimed at beating others, and *cooperative* efforts. Little empirical research is available on the effectiveness of competition as a motivator, but what little there is suggests that it is not very effective compared to individual and cooperative effort.[9] Nor is this really surprising since there are many reasons to strive for the goods that competition can make possible. Consider again the example of cancer research. Other, entirely noncompetitive reasons suffice to motivate individuals to discover a cure as speedily as possible: cancer is killing people, possibly including the researchers themselves. Moreover, a *cooperative* effort among cancer researchers has significant advantages over competition (and over individual effort). The aim of a cooperative effort is to find the cure as speedily as possible, and that calls for division of labor and sharing of resources and results. But the aim of competitors is to *win*, thus motivating them to withhold data and resources from their rivals.[10]

My quarrel (to pick up the thread) is not with all competitive desires, only necessarily nonuniversalizable ones. Elsewhere[11] I have tried to show that competitive desires in general are overrated. I will not repeat that discussion here. What still requires an explanation is why I condemn necessarily but not contingently nonuniversalizable desires even though both may generate aggression. There are two reasons. First, with a change of circumstances perfectly unobjectionable desires may be transformed into contingently nonuniversalizable ones, and we cannot possibly expect people to abandon them when the transformation occurs. Even the desire for enough food to keep body and soul together, we have seen, cannot be satisfied by everyone in every set of circumstances, yet expecting people to abandon the desire for food would be absurd. Second, while pursuing contingently nonuniversalizable desires *may* generate competition and hence aggression, it need not. Instead, we expect people to minimize their aggression out of a sense of morality, by *cooperatively* pursuing their aims. For example, if by certain cooperative arrangements more food could be produced quickly enough to keep all alive, it is possible to avoid aggression.

Compare necessarily nonuniversalizable desires such as being the best boxer. Even in serious boxing events the level of aggression may be reduced through the provision of rules of "fair play," just as rules of war (like the Geneva Convention) may reduce suffering. But war fought by the rules is still aggression, not cooperation, and so is any other contest "played" by the rules. Fighting contests by rules that limit aggression may be morally preferable to fighting them by no rules at all. Yet the aggression generated by the desire to be the best boxer can be prevented by a solution not available in the case of aggression generated by the desire for enough food: *abandon the desire*. My thesis is that all necessarily nonuniversalizable competitive desires are prima facie objectionable because they generate easily avoidable aggression. Still, my position is viable only if aggression carries a presumption against it. The nature of that presumption is complicated.

The wrong of aggression can be put as follows: it is prima facie objectionable due to the fact that it consists in acts that cause a good deal of intended or at least anticipatable misfortune for others. An act just is not aggression unless its upshot for others is dire. Moreover, aggression is the definitive case of the use of people as mere means.

Even aggression is permissible, however, in certain circumstances. The most obvious case is that in which it is the only means to avoid a much greater, catastrophic, wrong. There is a second case, I think, though a great number of readers will disagree with me on this point, and I lack the space to make a decent argument. I am inclined to say that it is permissible for people to do anything whatever to one another so long as all those affected give their fully informed consent. Others may even cause us misfortune and use us as a mere means if we give our fully informed *consent* to that treatment, say in exchange for permission to treat them likewise. Some (like Robert Nozick[12]) might not think that we *can* consent to being used since our consent is the mark of and a sufficient condition for our not being treated as mere means. But consenting to the acts of aggressors does not transform them into ones by which we are not treated as mere means. A serial killer who takes your life for kicks has used you even if you gave him permission (you were, we might suppose, on your way home to cut your own throat anyway). At any rate, if consent *were* a sufficient condition for not treating someone as a mere means, then it would be even more difficult for anyone to reject my assumption that aggression may be legitimized through consent.

On the basis of this assumption, I conclude that even contests that can be quite deadly may be mor-

ally permissible if conducted with the informed consent of all participants. Olympians scheduled for dangerous events (such as boxing matches) want the contests to occur; they do not want to lose out in them, but they prefer that to not being able to participate at all. It is, of course, extremely difficult to determine in any actual situation whether or not individuals have really given their fully informed consent to being caused misfortune. Nonetheless it is easy to imagine people who know exactly what they are doing in entering dangerous contests. None of them wants to be done a misfortune, but each is willing to exchange the requisite permission for license to treat the others likewise.

Nonetheless, consent legitimates aggression only in circumstances in which the aggression can be *contained*, so that it will not result in harm to bystanders: duels in crowded supermarkets would subject shoppers to stray bullets. The containment problem is especially dangerous when entire nations are involved.

In part, the case against the pursuit of nonuniversalizable competitive values and positional goods is that such pursuits involve us in aggression, so that the strong presumption against the latter is inherited by the former. But even if such pursuits did not generate aggression there would be a case against them. One charge is that to flourish at the expense of others, which is what those who attribute importance to seeking nonuniversalizable competitive properties propose to do, is to treat others as a mere means. A second charge might be made by John Rawls. His apparatus can be used to provide an argument against the attribution by a just society of importance to excelling and other nonuniversalizable competitive properties.[13] The parties in the original position would certainly avoid a conception of justice that attributes much importance to excelling, since they would be concerned to avoid the resulting situation in which those who do not excel are adjudged by society to be second-class citizens, which would be fatal to their sense of self-esteem, perhaps the most important primary good according to Rawls. If a just society is one in which those who excel are due more respect than those who do not, then it will be impossible for everyone to form a commitment to justice so conceived, which is tantamount to saying that there's no solution to the problem of justice, the problem of providing the terms of association for everyone in society. (Of course, the parties would have no choice but to attribute importance to excelling if it were simply a matter of psychological fact that everyone considers it important. The above argument—as well as Rawls' own defense of his difference principle—rests on the assumption that the value that is placed on excelling is socially inculcated.)

The moral objections to pursuing nonuniversalizable competitive values are not the only objections. There is also the fact that absurdities result when we consider the satisfaction of positional and nonuniversalizable competitive desires to be of great importance, or value things through the conatus of such desires. The main point is that attributing intrinsic value to excelling and its kin is to commit ourselves to the absurdity that no matter how rich our lives are in noncompetitive goods, they still lack something of great importance if the lives of others *share the same* noncompetitive goods. To say that life can be worthwhile only if we excel is worse: it is to say that no matter how rich our lives are in noncompetitive goods, they are worthless if those goods are shared by all. I elaborate on this point elsewhere, concluding that such properties contribute virtually nothing of significance to our lives.[14]

Even the positional goods that are inspired by competitive pursuits involve us in an absurdity: the collective effort to acquire positional goods requires great effort yet gets us nowhere. Each move we make toward achieving these goods helps undermine their value for others, and *vice versa*, so that the closer we each come to achieving the goods, the closer they come to being without any value at all. Practices that get us nowhere, Sisyphian tasks like running on a treadmill, are obviously inane. Far better off are those who never allow themselves to be concerned about positional goods in the first place. Not only can such people avoid the frustration just described, they can also take advantage of the tendency of competitivists to pay noncompetitivists for the privilege of status. This tendency is described in detail by Robert Frank in his engaging book *Choosing the Right Pond*.[15] The main idea is that since not everyone can acquire items if they are to retain positional value, people who get them can be expected to pay those who do not for the privilege.

Note finally that in any competition losing is not the only way to fail to win; tying is another way, and preferable by any competitor to losing. Moreover, since only the *few* may win, then among com-

petitivists the majority is likely to prefer equality to the only alternative for them, namely losing. They cannot win, but at least they may avoid losing.

So there is a sense in which egalitarianism is the majority's unstable and strained solution to the competitivists' irrational struggle to outdo each other; instead of dissolving the competition by abandoning competitive values, each agrees to be frustrated in exchange for the frustration of the others. It seems likely that the interest in egalitarianism *vis-à-vis* goods other than the liberties is based on competitivist values, so that if we rejected competitivism, we would reject all reason to embrace egalitarianism. Unless I am interested specifically in how I or my holdings (goods, wealth) stack up against others, why would I be concerned about the fact that *you* have more than I? So long as I have enough, why should it matter to me that you have *more*? I hasten to add that a rejection of competitivism does not support indifference to the plight of others. The point is that it is one thing to be concerned about others on the grounds that given the intrinsic features of their situation (e.g., the fact that they are starving) it is clear that they need help, and it is quite another to be concerned about others because of their *relative standing* to us. The latter, I suggest, is a matter of indifference. Thus while my position alleviates the concern that there are (in Hirsch's phrase) "social limits to growth," it also helps undermine the belief that justice calls for equality or for other arrangements concerning the relative standing of others.[16]

Trinity University

Received July 10, 1989

NOTES

1. A small sample: M. Midgley, *Wickedness* (London: Routledge & Kegan Paul, 1984); S. Freud, *Beyond the Pleasure Principle*, J. Strachey, trans. (New York: W. W. Norton Company, 1961); K. Lorenz, *On Aggression* (New York: Harcourt, Brace and World, 1966); I. Eibl-Eibesfeldt, *Love and Hate* (New York: Holt, Rinehart and Winston, 1971); A. Montagu, *The Nature of Aggression* (New York: Oxford Press, 1976); E. O. Wilson, *On Human Nature* (New York: Bantam Books, 1978); and P. Kitcher, *Vaulting Ambition* (Cambridge, MA: MIT Press, 1985).

2. See *Superiority and Social Interest*, H. and R. Ansbacher, eds. (New York: W. W. Norton, 1979), p. 31.

3. R. Wasserstrom, for example, says this in his influential paper, "On the Morality of War: A Preliminary Inquiry," *Stanford Law Review*, vol. 21 (1969), pp. 1627-1656.

4. This thesis is developed by R. Carneiro in "A Theory of Origin of the State," *Science*, vol. 169 (1970), pp. 733-38.

5. In *Social Limits to Growth* (Cambridge: Harvard University Press, 1976), pp. 20-23, 27.

For commentary on Hirsch's views, see A. Ellis and K. Kumar (eds.), *Dilemmas of Liberal Democracies: Studies in Fred Hirsch's "Social Limits to Growth"* (London: Tavistock Publications, 1983); M. Hollis, "Positional Goods," *Philosophy and Practice*, A. Phillips Griffiths, ed., Royal Institute of Philosophy Lecture Series, No. 18 (Cambridge University Press, 1985); John Robertson, "Honour and the Good Life," unpublished Pacific APA paper, and R. Frank, *Choosing the Right Pond* (New York: Oxford University Press, 1985).

6. I shall also use the term positional *desire* to refer to the attitude whereby we regard something as a positional good. We may add that a desire for something *X* is *positively* positional for me just in case it is coupled with the disposition to value *X* more to the extent that less people other than me acquire *X*; it is *negatively* positional for me just in case it is coupled with the disposition to value *X* less to the extent that more people other than me acquire *X*.

7. Contrast R. Nozick's approach in *Anarchy, State and Utopia* (Cambridge, MA: Basic Books, 1974), pp. 239ff.

8. But see the qualifications in my "Competing for the Good Life," *American Philosophical Quarterly*, vol. 23 (1986), pp. 167-77.

9. A great deal of empirical data concerning the effectiveness of competition in an educational setting is surveyed in D. W. Johnson, G. Maruyama, R. Johnson, D. Nelson, and L. Skon, "Effects of Cooperative, Competitive, and Individualistic Goal Structures on Achievement: A Meta-Analysis," *Psychological Bulletin*, vol. 89 (1981), pp. 47-62. See the engaging discussion of these data by A. Kohn in his *No Contest: The Case Against Competition* (Boston: Houghton Mifflin Company, 1986).

10. My position on competitive values appears to undergo attack by John Kekes in "What Makes Lives Good?" *Philosophy and Phenomenological Research*, vol. 48 (1988), where he claims that status and prestige are perfectly reasonable goods for us to pursue. However, I find no argument for this, aside from the obvious point that they have instrumental value.

11. In "Competing for the Good Life," *op. cit.*

12. Robert Nozick, *Anarchy, State and Utopia*, *op. cit.*

13. John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard Press, 1981).

14. In "Competing for the Good Life," *op. cit.*

15. *op. cit.*

16. I thank Frances Berenson, Curtis Brown, Herbert Fingarette, Robert Frank, Daniel Kading, Susan Luper-Foy, John Robertson and Mark Williamson for many helpful criticisms and comments. I also thank Trinity University for providing a stipend which made work on this essay possible.

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 27, Number 3, July 1990

ON THE ALLEGED METHODOLOGICAL INFIRMITY OF ETHICS

Michele M. Moody-Adams

TWO deceptively simple claims, accepted by some as unchallengeable givens, have shaped much twentieth century philosophical discussion of the nature of moral theory and moral argument. The first claim is that moral discourse is ultimately non-rational because it can consistently generate disputes which appear resistant to solution. I call this the "non-rationality of intractable disagreement" thesis. The second claim is that because theory-construction in ethics cannot be shown to be substantially analogous to theory-construction in science, it must rely on methods which are rationally deficient. Even some thinkers who reject foundationalist epistemology argue that science has sufficiently strong links to observation to render scientific claims susceptible of a correspondence theory of truth, while ethics—to adopt a phrase of Quine's—is "methodologically infirm" as compared with science.[1] Thus I call this the "methodological infirmity" thesis.

My aim in this paper is to defend the rationality of ethical discourse, and the vigor of the methods of ethics, by showing that these two theses—of the non-rationality of intractable disagreement and of methodological infirmity—rest on serious misconceptions about the nature of theory construction and argument in ethics. The methodological infirmity thesis rests on the presupposition that the methods of ethics and the methods of science can be, and ought to be, compared. I argue that the connection between ethical theory and experience is simply different *in kind* from that between scientific theory and experience—so different, that any demand to show even a weak analogy between the methods of ethics and those of science must be fundamentally misconceived. The thesis of the non-rationality of intractable ethical disagreement rests on an equally untenable assumption: that the purpose of argument in ethics is always, and could only be, to secure agreement. I argue, on the contrary, that moral reflection sometimes accomplishes its purpose merely by stimulating individual or collective self-scrutiny. Indeed, certain kinds of ethical disagreements are central to the rationality of ethics—even to rationality itself—because even intractable ethical disagreements are important catalysts for self-examination. Much recent philosophical discussion of moral theory and moral argument has overlooked the important connection between moral reflection and self-scrutiny. I will show how this connection makes unintelligible any expectation that ethical theories can be "tested" in the manner of scientific theories.

Philosophical formulations of doubts about the methodological adequacy of ethics derive much of their force from claims about moral disagreement. I thus consider three influential arguments arising from reflection on this phenomenon: arguments made by A. J. Ayer, C. L. Stevenson, and, in particular, W. V. O. Quine. Quine presents the most forceful version of the infirmity thesis, but his view is most intelligible in the context of the emotivism which inspired it. I then consider some strategies for responding to these arguments—in particular, to Quine's argument. One important response to Quine's position, offered by Owen Flanagan, attempts to show that at least some standards of argument and "testability" in science can be applied to reflection in ethics.[2] But any such strategy is bound to be unsuccessful because it rests on some of the same misconceptions as the views it is intended to rebut. Instead, a challenge to the infirmity thesis must avoid these misconceptions by restoring the connection between moral reflection and self-examination to its rightful place in moral philosophy.

I

Ayer's early reflection on moral disagreement led him to insist that, often, what looked like argument about some ethical matter could not really be an

argument at all.[3] For example, when I say that stealing money is wrong and another person insists that there is nothing wrong with stealing, that person, according to Ayer, cannot "strictly speaking" contradict me. Neither of us is "asserting a genuine proposition," but instead giving vent to certain feelings. The judgment "Stealing money is wrong," as a "pure" judgment of moral value, is "outside the scope of argument." Ayer did hold that some moral judgments may not be purely about values; dispute over such judgments might reduce to "argument about a question of logic" or about "an empirical matter of fact." Argument about *purely* ethical judgments, however, is impossible: "[i]t is because arguments fail us when we come to deal with pure questions of value, as distinct from questions of fact, that we finally resort to mere abuse."[4] Ayer gradually retreated from the most counterintuitive features of his early statement of emotivism, but he continued to treat the intractability of some ethical disagreements as evidence that ethics is not fully rational.

C. L. Stevenson, Ayer's most influential emotivist successor, thought that ethical arguments did occur and that difficult ethical disagreements could sometimes be resolved, yet he insisted that their resolution would typically require the use of "non-rational methods." He based this claim on a distinction between two kinds of disagreement: disagreements in "science, history, and their counterparts in everyday life" involve an opposition that is "primarily of beliefs," while disagreements in ethics involve primarily an "opposition of attitudes."[5] He never claimed that every dispute in science or history is exclusively a disagreement in belief, nor that every ethical argument is exclusively a disagreement in attitudes. Moreover, he did suggest that some ethical disputes might be *primarily* disagreements about beliefs—as when divergent evaluations of an action rest on ignorance of the fact that the action is actually a means to some end which all parties to the dispute share. Since we can try to adduce reasons to revise the offending beliefs, such disputes are susceptible of resolution by means of "reasoning and inquiry." Yet "if any ethical dispute is *not* rooted in disagreement in belief, then no reasoned solution of any sort is possible." Such disputes can be resolved only by resort to "non-rational methods": persuasion, exhortation, various forms of public demonstration and display, and material rewards and punishments.[6]

But Stevenson's view rests on some implausible assumptions. His account of kinds of disagreement implicitly presupposes that any realm of discourse in which argument and theory construction fail to proceed *solely* by means of the methods of science must be ultimately non-rational, a matter of disagreement in non-rational attitudes rather than in (potentially rational) beliefs. This presupposition, if correct, would require Stevenson to treat the discipline of history, for example, as ultimately non-rational. This striking conclusion follows from Stevenson's implausible view of the methods of history—which he treats as precisely analogous to the methods of science.

Generalizations drawn from science will surely figure in historical reconstructions, yet the reconstruction of complex eras or events is not simply a matter of applying scientific methods to the past. For instance, the kinds of justificatory methods available to a scientist who cannot appeal to experience alone are generally unavailable to the historian. A scientist may sometimes convince us to accept a new theory by reference to its simplicity, its elegance, or its conservation of current beliefs. But we have no reason to assume that the simplest, most elegant, or most conservative historical reconstruction is most likely to be correct. It is largely because of this difference between history and science that historical reconstructions may generate disagreements as intractable as any ethical disagreement. Of course, were historians to refrain from making any claims not reducible to questions of logic, to "the facts" independent of any "attitudes" about what happened, or to implications of scientific generalizations, intractable disagreement in history might be rare. But what kind of history might such constraints produce? The possibility of intractable disagreement in history shows that the relation between historical theory and experience is different in kind from that between scientific theory and experience; so, too, are the methods of the two disciplines. Stevenson's failure to recognize this difference, or to see that a discipline not structured on the model of science can be rational, shows that the monistic conception of rational methodology embodied in his account rests on an implausible picture of the notions of argument and justification.

II

Quine offers a more compelling formulation of

doubts about the rational adequacy of the methods of ethics. He first reminds us that moral disagreement may take the form of cross-cultural disagreement, disagreement between persons in the same culture, even one person's private uncertainty about a single issue. When such disagreement occurs, he contends,

> one regrets the methodological infirmity of ethics as compared with science. The empirical foothold of scientific theory is in the predicted observable event; that of a moral code is in the observable moral act. But whereas we can test a prediction against the independent course of observable nature, we can judge the morality of an act only by our moral standards themselves. Science, thanks to its links with observation, retains some title to a correspondence theory of truth; but a coherence theory is evidently the lot of ethics.[7]

Some who would defend the vigor of the methods of ethics take this passage as a challenge to show that normative ethical theories are answerable to experience in the same manner as scientific theories. Quine thinks that the challenge cannot be met, although the failure to meet it need not force us to accept all actions or ways of life as equally valuable:

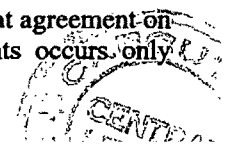
> Even in the extreme case where disagreement extends irreducibly to ultimate moral ends, the proper counsel is not one of pluralistic tolerance. ...We can still call the good good and the bad bad, and hope with Stevenson that these epithets may work their emotive weal.[8]

But this response doesn't take us very far beyond the standard emotivist view that ethical discourse is ultimately non-rational. Thus it makes sense to examine Quine's account of the rational adequacy of the methods of science.

A scientific theory, in Quine's view, is part of a vast network of sentences linked by means of inferential connections.[9] On the "periphery" of this network are certain sentences—"observation sentences"—which have a close link to experience. The methodological vigor of a theory is a function of the extent to which that theory is capable of implying observation sentences. But Quine's physicalism requires a very spare picture of the phenomenon of language, even the language of scientific theory. Language, for Quine, consists of complex patterns of sounds and utterances; these utterances, and our dispositions to produce them, are the primary data for any study of language. Quine begins with the notion of "stimulus meaning": the disposition of a speaker to assent to, or dissent from, a sentence in response to certain sensory stimulation. Observation sentences are those sentences which command the same verdict from a group of speakers when the following conditions are met: (1) the sensory receptors of all speakers in the group are stimulated in roughly the same way, and (2) the disposition to assent to, or dissent from, the sentence hinges on "collateral information" that is publicly shared. Observation sentences thus might be defined as sentences for which the stimulus meanings for different speakers of a language tend to coincide. There will be firm *agreement* on the truth-values of such sentences among sincere and well-placed observers who speak the same language. A sentence such as "It's raining" is observational in the required sense. Even the disposition to assent to or dissent from a sentence like "That's a rabbit," can hinge on *publicly shared* collateral information, and thus possesses a degree of "observationality."[10]

Quine acknowledges that there are degrees of observationality; even scientific theories contain sentences for which the disposition to assent or dissent depends upon extremely complex inferential links with other sentences. Yet he believes that scientific theories always imply at least some observation sentences, whereas ethical theories do not, and that science is thus answerable to experience in a way in which ethics cannot be. To be sure, on Quine's holism, no *single* sentence of a scientific theory could have a "separable bundle" of testable consequence—unless we make "one long conjunctive sentence" of a whole theory.[11] Yet a scientific theory, or even a reasonably inclusive portion of such a theory, will imply a series of conditional statements of the following sort: "If observable conditions of a certain kind obtain, then a certain kind of observable event will occur." Both the antecedent and the consequent of such conditionals will be observational in the required sense. A scientist can thus test a portion of her theory by treating it as vulnerable to the test of observation conditionals—while treating the rest of the theory as firm, at least for the time being.[12]

Quine's assertion of the methodological infirmity of ethics is ultimately a claim that no sentence implied by an ethical theory could be an observation sentence—even though, on occasion, a sentence such as "That's shameful" might elicit the same response from a group of well-placed observers. Implicit in this claim is the belief that agreement on the truth-value of ethical statements occurs only



when speakers already accept the theory, or that portion of the theory, which the alleged observation sentence was to help confirm. In other words, Quine thinks that we will assent to "That's shameful" only if we already accept as authoritative the moral standards which require such a verdict. Quine must admit, of course, that the disposition to assent to a sentence like "That's a rabbit" hinges on information which goes well beyond present stimulation of the sensory receptors. But he can also insist that an observer need not accept any portion of the very theory being tested (say, a genetic theory which implies an observation conditional containing the consequent "That's a rabbit") in order to be disposed to assent to "That's a rabbit." That same observer's disposition to assent to "That's shameful," in contrast, will hinge upon her acceptance of the ethical theory at issue; the "information" which goes beyond present stimulation of the sensory receptors includes the theory being tested.[13]

Quine's argument is quite compelling even if one believes—as I do—that agreement on the truth or falsity of statements in both ethics *and* science presupposes not only a context of shared beliefs about how to structure experience, but also consensus on the criteria of good arguments, and even shared values or ends. That is, even if one holds that some deep "agreement in judgments" underwrites surface agreement on the truth of statements in science as well as in ethics, Quine seems to have identified an important difference between discourse in ethics and discourse in science. Sincere and well-placed observers who share a language (who can agree about the truth of sentences concerning rabbits and electrons, and sometimes even black holes) can disagree about matters as fundamental as the moral status of abortion, the moral acceptability of using violence to promote desirable political ends, or the moral acceptability of consuming meat. Members of a single linguistic community can and do disagree about the truth-value of sentences like "killing animals for human consumption is shameful" (or, even, "It is wrong of that man to kill that rabbit")—even in circumstances where all would agree on the truth of "That's a rabbit."[14]

III

A response to Quine might begin by rejecting his assumption that the empirical foothold of ethical theory is in the observable moral act, in order to sidestep the problem of looking for an ethical counterpart to the observation sentences implied by scientific theories. Owen Flanagan has tried just such a strategy. Flanagan points out that Quinean epistemology naturalized could not be a purely descriptive endeavor, since it

> retains a factual side comprised by learning theory and evolutionary theory and a normative side which consists of the sorting of successful belief strategies from unsuccessful ones—for example, the scientific method from gambling. Practice...is the sorting device which generates the normative epistemologist's principles.[15]

The naturalistic epistemologist must tell us which rules for acquiring beliefs are successful belief-gathering strategies, and the task of doing so is a normative enterprise. Moreover, the empirical foothold of this enterprise is in the ongoing practice of theory-formation. Thus a normative ethics modeled on normative epistemology would have its empirical foothold not in "observable acts" but in ongoing moral practice.

Flanagan notes that Quine offers the outlines of a descriptive ethics which parallels the structure of his descriptive epistemology. For instance, Quine explains our capacity to acquire moral values by reference to an innate tendency to order experiences along a "valuation axis."[16] He suggests that features of the central nervous system determine the ordering of certain experiences along this axis, so as to produce much of the cross-cultural agreement in morality. Still further, he posits a capacity to "transmute means to ends"—by means of which the ordering of sensory episodes along the valuation axis is revised and refined—to explain the possibility of moral education. Flanagan believes that since Quine's naturalism can license such a descriptive ethics, that naturalism should allow for a normative ethics, modeled on normative epistemology, that is also compatible with science. This naturalized normative ethics would appeal to practice to identify "successful" moral strategies. For instance, a naturalistic ethical theorist might appeal to the practices of groups who value the quality of truthfulness in persons, to determine that norms requiring truthfulness are "successful" moral strategies.

To be sure, the sorting of successful belief strategies is part of Quinean epistemology naturalized. Quine takes reflection on successful hypothesis-for-

mation in science to point to five "virtues" that plausible hypotheses tend to possess, and to yield "guides to the framing of hypotheses" intended to be normative for the practicing scientist.[17] The naturalistic epistemologist need not "jettison the normative and settle for the indiscriminate description of ongoing procedures."[18] Yet Quine thinks that normative epistemology has one characteristic which normative ethics could never possess:

> normative epistemology is a branch of engineering. It is the technology of truth-seeking, or in a more cautiously epistemological term, prediction. Like any technology, it makes free use of whatever scientific findings may suit its purpose. ...There is no question here of ultimate value, as in morals; it is a matter of efficacy for an ulterior end, truth or prediction. The normative, here, as elsewhere in engineering, becomes descriptive when the terminal parameter is expressed. We could say the same of morality if we could view it as aimed at reward in heaven.[19]

Quine thinks that the normative epistemologist's claims are tied to an unassailable criterion of success: true, or reliable, predication. No ethical theorist, in his view, even in observing the ongoing practice of "moral strategies," could find an equally unassailable criterion to which to appeal.

Quine's formulation of this obvious contrast between normative ethics and normative epistemology is, however, vulnerable to an important objection. For in order to accept any account of successful belief-gathering strategies, we must first accept the criterion of success presupposed by that account. As Hume insisted, one reason we value theories which generate true predictions is because of the role such predictions play in helping us to promote our aims and interests.[20] Hume's point can even be expressed in language which Quine might want to confine to the explanation of moral education. We might say that it is by a process of the "transmutation of means to ends" that we come to value the end of true predictions, much as we once valued only the benefits to which such predictions were just a means. A genealogy of the epistemologist's "ulterior end" thus shows that his criterion of epistemological success is as much an "ultimate value" as is any end appealed to in ethics. Of course, Quine would nonetheless insist that there is unanimity, or at least sufficiently general agreement, on the value of reliable predictions in science, and that there is no such general agreement on a criterion of success in ethics—and on this, he is surely right. The notion of the ongoing practice of "successful moral strategies" cannot play the role Flanagan wants it to play: there simply is no sufficiently general agreement on what constitutes a "successful" moral strategy.

IV

In trying to align the methods of ethics with those of science, Flanagan also calls attention to the role of thought-experiments in ethics and in science. He notes that scientists engage in "imaginary practice," testing hypotheses out in imagination, rather than in the world. Ethical theorists likewise construct imaginary possible worlds, in which various moral conceptions "figure in the overall economy of social life," and they rely upon discourse about such worlds to generate support for an ethical theory.[21] The notion of "imaginary practice," Flanagan believes, allows one to bypass worry about the empirical rock-bottom of ethical theory.

Theory construction in science obviously relies on what Flanagan calls "imaginary practice." Indeed, it is sometimes unclear what would count as a suitable test of the empirical significance of theories constructed in this manner. This is especially true in physics, where the strongest initial support for a thought-experiment is frequently not predictive success but mathematical consistency. Yet mathematical consistency is taken to support hypotheses in physics because it has been a reliable guide to many past advances. Even where the *immediate* goal of scientific theory-construction is not successful prediction, the *ultimate* goal is an explanatory theory compatible with theories that *can* generate reliable predictions. The failure to produce such a theory does constrain extrapolation in science. But when we turn to the ethical theorist's use of imaginary practice, it is not easy to specify what constrains extrapolation in ethics. The ethical theorist's imaginary practice typically appeals, somehow, to our intuitions. The goal of the thought-experiment is to match, to "capture," or to impose a more coherent structure upon those intuitions. But the failure of an ethical theory to match our intuitions cannot constrain extrapolation in ethics in the way failure of prediction (or the failure of a mathematically elegant explanation to cohere with reliable predictive hypotheses) constrains extrapolation in science.

This disjunction is borne out by two of the most influential thought-experiments in recent moral philosophy: those in John Rawls' *A Theory of Justice* and Robert Nozick's *Anarchy State and Utopia.* Each theorist claims to start from what is ostensibly the same intuition: that it is wrong to treat persons merely as means to the promotion of some social purpose. Yet the moral conceptions designed to capture this intuition are obviously very different. In Rawls' just society, social institutions, subject to the priority of the principle of equal liberty and the fair equality of opportunity, will be arranged so that inequalities in wealth and power benefit the least well-off representative person. Thus his conception of justice may require taxation to finance transfer payments. On Nozick's account, the only legitimate state is a "night-watchman" writ large, which protects citizens against force, fraud, and theft and ensures the performance of contracts. For Nozick, taxation to fund anything more than this minimal state violates the prohibition against treating persons merely as means to some social end.

No ethical theorist would argue that appeal to intuitions *alone* could settle the dispute between these very different moral conceptions (nor is Flanagan committed to saying anything of the kind). But if one is to complete Flanagan's attempted analogy between thought experiments in ethics and in science, one would have to add *something* to intuitions—and it is not clear what that could be. It is especially unclear how to complete the analogy if ethical theories cannot be shown to be ultimately confirmed or disconfirmed by observation—as science presumably can. Flanagan reminds us that theory-construction in science is not constrained by observational feedback alone, but by observation along with considerations such as consistency with, and conservation of, existing beliefs, as well as simplicity and generality of explanation.[22] Yet consistency, conservatism, simplicity, and generality have proven to be reliable—successful—guides to theory-construction in science. Once again, the fact of general agreement on a criterion of scientific success reveals a fundamental difference between the methods of ethics and the methods of science.

V

Once one grasps the real empirical foothold of ethical theory, and the distinctive relation between ethical theory and experience, it will become clearer why the methods of ethics cannot be analogous to the methods of science—and why, moreover, we wouldn't want them to be. An ethical theory's foothold in experience is not Quine's "observable act," nor Flanagan's ongoing practice of "moral strategies," nor even (except indirectly) the moral intuitions which figure in the construction of ethical thought-experiments. Rather, the empirical foothold of moral theory, as of the garden-variety moral argument, is in the *self-understanding* of those addressed by the theory, or of those embroiled in the debate. This is the (perhaps unexpected) kernel of truth embodied in Ayer's sardonic comment that intractable moral disputes may conclude with resort to "mere abuse." Any moral stance—whether a sophisticated philosophical theory, a loosely connected set of intuitions, even an incoherent jumble of judgments and attitudes—will be bound up with particular visions of the sort of persons who accept or reject such positions or intuitions, or who make such judgments and hold such attitudes.

I intend expressions such as "self-understanding" and "self-conception," which I shall treat as roughly synonymous, to be broadly construed. One's self-understanding includes not only beliefs about oneself, but also about one's place in the natural (and possibly the supernatural) world, and especially about one's relationships to other selves—about one's place in the social world. Moreover, a self-conception is a product of both social and individual influences. To be sure, one's self-conception is partly shaped by the peculiarities of one's formative experiences. But formative experiences include *social* experience; some components of a self-conception owe their origin to something independent of the individual person and her private experiences.

Further, one's self-conception, by its very nature, can always be revised. In fact, one function of ethical theory, as of moral reflection in general, is to encourage the sort of self-scrutiny which may lead us to see ourselves, our relations to others, and our place in the world in a different way. Plato's early *Dialogues* provide the purest examples of this conception of moral reflection. When, for instance, we wish that Euthyphro would stay with Socrates to begin his discussion of "piety" all over again we discover one of the essential characteristics of moral reflection. Moral reflection should serve to stimulate the kind of self-scrutiny which Euthyphro was reluctant to

undertake. Of course, we often want moral reflection to effect a change in practice as well as in self-understandings. But moral reflection can generate reform in practice only if it first encourages self-scrutiny. The success of American civil disobedience in the 1950's and 1960's was a function of its success in encouraging the collective self-scrutiny of a whole culture—success which reveals the shortcomings of Stevenson's claim that "public demonstration and display" are necessarily non-rational. Of course moral reflection is not the only method for generating self-scrutiny: violence and torture sometimes seem to have this effect. But this simply shows that we cannot defend some practice simply because it may tend to encourage self-examination. Further, self-scrutiny will not always produce radical change in self-understandings, nor will it always generate new moral theory or some new stance on a moral issue. But there is no reason to think that moral reflection is valuable only if it yields such results—understanding the presuppositions of a moral stance makes it easier to decide difficult issues in the future.

My view requires no metaphysics of the self. There may be no knowable entity over and above the various self-conceptions to which any given person would assent over a lifetime, nor need we be able to show that there is such an entity in order to do moral theory. We must be able to show only that persons possess some conceptions, however vague, of their place in the natural, social, and possibly supernatural worlds, and that such conceptions can be revised. Indeed, the revisability of self-understandings is a condition of the possibility of moral theory.

Critics of moral philosophy who overlook the central role which self-scrutiny plays in moral reflection are bound to misunderstand moral argument and normative moral theory. The most common misunderstanding is embodied in the expectation that an ethical theory's foothold in experience should be capable of "confirming" or "disconfirming" that theory. The relation between ethical theory and experience makes any such expectation unintelligible: ideally, engaging in moral reflection will change the very structure of the relevant "experience." Equally problematic, many critics do not see that the basis for the possibility of moral agreement is some degree of overlap in the content of self-conceptions. A person who believes that the product of conception is infused with an immortal soul at the *moment* of conception will very likely disagree about the moral status of abortion with one who does not see herself in this way. Such a failure to reach agreement does not show the non-rationality of moral discourse, but how difficult it is to convince another to see the world and her place in it in a new way.

The Rawlsian democratic egalitarian fails to convince the Nozickean libertarian that the part of Rawls' conception requiring the financing of transfer payments—the Difference Principle portion of Rawls' second principle—is defensible. Rawls' Difference Principle is addressed to persons who can accept a very distinctive vision of themselves. Rawls must be able to convince us both that there is no morally relevant sense in which we deserve our fortunes in the "natural lottery," and that any gain arising out of the arbitrary good fortune of those who do well in that lottery must be compatible with an institutional expression of concern for the expectations of those who have arbitrarily lost out.[23] The libertarian will concede the first revision in his self-conception, but not the second. He will insist that concern for the expectations of others cannot be institutionalized as Rawls' theory requires without violating the prohibition against treating persons merely as means. But the failure of Rawls' theory to convince the libertarian is not evidence of the non-rationality of moral discourse. Rawls simply fails to convince the libertarian to revise his self-conception in the manner required to secure agreement on the notion that the Difference Principle is part of a defensible conception of justice.

Once we understand that the empirical foothold of ethical theory is the self-understanding of those to whom the theory is addressed, the most important characteristics of ethical method become obvious. In particular, it should be clear that the relevant aspects of the self-conceptions of those to whom ethical theory is addressed are accessible to the ethical theorist only by means of the moral intuitions to which they would assent at a given point in time. Contemporary ethical theorists who announce their intentions to begin with intuitions simply do explicitly what ethical theorists have, implicitly, always done. Kant, for instance, took himself to have revealed the rational structure presupposed by central elements of the ordinary moral consciousness. He insisted that, like Socrates, he was making human reason "attend to its own principle," and he relished one critic's observation that the *Groundwork* had told us "nothing new" about morality.[24] Of course, Kant

also believed that the rational structure revealed in the *Groundwork* could be shown to be necessary and unconditionally valid—valid independently of the fact that we could agree to that structure. Many contemporary ethical theorists who proceed from intuitions would, in contrast, defend roughly "constructivist" accounts of the content of morality. On such accounts, the objectivity of moral claims is somehow constituted by those who accept those claims, and the independent validity of those claims is neither affirmed nor denied.[25] But ethical theory is pointless if it is not addressed to persons whose behavior it aims to influence. It *must* start from the "inside": from the pre-theoretical deliverances of the moral consciousness of those to whom the theory is addressed.

Critics of my claim will ask what such a view presupposes about the status of the intuitions which precede theory-construction in ethics, and what we are to make of any theory which begins from intuitions. According to one objection, if we construe moral intuitions as deliverances of an innate and essentially trustworthy moral capacity, and then take the task of ethical theory to be that of bringing rational coherence to the uses we make of this capacity, the resulting theory is just an elaborate exercise in moral psychology. On Harman's version of this objection, the theory will be a study in commonsense ethics, and such a study is no more an ethical theory—"the theory of rightness and wrongness"—than a study of commonsense physics is a study in physics.[26] Critics may argue that the only plausible alternative construal of intuitions fares no better. If moral intuitions are not deliverances of a natural moral competence, it would seem that they are mere accidents of upbringing, or "prejudices." Understood in this way, intuitions appear to have *no* special standing at all, and thus cannot guide the construction of ethical theory.

My response to these objections must be clear. To be sure, the judgments which those addressed by an ethical theory would be willing to make are the only vehicle through which the theorist might reach the self-conceptions underlying a moral conception. Yet I do not claim that moral intuitions alone can (or should) constrain ethical theory construction. Interest in normative theory is typically generated in circumstances which reveal current moral intuitions to be in some way inadequate, or even inconsistent with one another. It doesn't really matter which construal of intuitions we take, since the acceptance or rejection of current moral intuitions must be guided by considerations of consistency, coherence and conformity with a self-conception capable of surviving rational scrutiny.[27]

Rawls' conception of justice reveals even more fully why neither objection can undermine an ethical theory *simply* because it starts from the "inside." Rawls aims to present a set of principles which we can start as most successful in bringing coherence to our reflection about justice. Yet though Rawls begins with intuitions ("considered convictions"), he knows that the principles which he believes to best secure rational coherence *may*, in particular cases, yield verdicts which initially fail to match our considered convictions about justice. Thus he attempts to provide considerations in favor of revising, or even relinquishing, some of the convictions which the theory initially fails to match. If he succeeds, then our judgments can be brought into agreement with the verdict which the theory requires in particular cases. We thus arrive at an agreement between convictions and theory which Rawls calls "reflective equilibrium."

Because Rawls has claimed to provide a "description" of our sense of justice, it is important to note that the deliberation leading to reflective equilibrium may require revision, or even rejection, of some pre-theoretical convictions.[28] For Rawls' claim to describe our sense of justice might suggest that the view is, after all, moral psychology masquerading as ethical theory. But Rawls himself suggests a way to understand his enterprise which is ultimately more illuminating: moral philosophy, he argues, is the study "of principles which govern actions shaped by self-examination."[29] "Moral philosophy is Socratic," he continues, and "we may want to change our present considered judgments once their regulative principles are brought to light."[30] This appeal to the Socratic model reveals that Rawls wants to restore self-examination to the central place in moral reflection which it was rightfully assumed to have on the Socratic view. Rawls' theory is neither an exercise in moral psychology, nor an attempt to treat pre-theoretical convictions as though they could alone constrain extrapolation in ethics. Rather, it is an attempt, by means of the rational scrutiny of pre-theoretical convictions and of the self-conception which underlies those convictions, to secure agreement on principles which ought to shape the structure and functioning of social institutions.

Understood in this way, Rawls' method can serve as a model for method in ethical theory in general. Theory construction in ethics is fundamentally the attempt to secure agreement on the results of the kind of rational self-scrutiny that is aimed at guiding reflection and conduct. The results may consist of a supreme principle, or principles, for guiding reflection and action, or a general conception of the character traits and emotions which ought to determine the shape of reflection and action. But guides to reflection and conduct which deserve the label "ethical" must rely, if only implicitly, upon canons of rational self-scrutiny.[31] Even a non-constructivist conception of objectivity—which Rawls, of course, does not defend—must aim to secure agreement on a particular vision of the content of moral reflection. A normative ethical theory is, after all, intended to guide conduct.

Of course, the theorist who is sceptical about the cognitive status of moral claims will wonder whether ethical theory as I defend it is epistemologically respectable. Would I claim that securing agreement on the results of the self-scrutiny which shapes moral reflection could amount to a justification of the theory which results? This is precisely what I want to argue, fully aware of the likely objections.[32] Critics may charge that justification involves showing the objective correctness of a theory, and that objectivity is a function of some kind of correspondence with a reality independent of the theory. It may be asked how the notion of *collective agreement* on the self-understanding underlying a vision of how things ought to be could have anything to do with showing an ethical theory to be objective.

I want to argue that the notion of collective agreement is inseparable from the notion of justification—in any realm of human discourse. If it is epistemologically unsound to link the notion of collective agreement with the notion of justification, then the objectivity of science will be in question. Justification in science is possible only because a complex, if largely implicit, collective agreement underwrites the possibility of agreement on the correctness of (even low-level) statements about the physical world. Such agreement concerns the structure of experience, the criteria of good arguments, and even the virtues which physical theory ought to embody. To be sure, a large portion of this agreement is deeply embedded in our discourse about the physical world—so deeply embedded that it is guaranteed merely by our membership in linguistic communities. Thus it is often overlooked in discussions of justification in science. Nonetheless this complex background agreement is what makes such justification possible. I do not want to deny that there is a difference between facts and values, and, more crucially, between discourse about facts and discourse about values. Little, if any, of the background consensus necessary to secure substantive agreement on moral claims is as deeply embedded in our discourse about the world as is the background agreement relevant to science. Further, the kind of background consensus capable of underwriting substantive moral agreement tends to be bound by more narrowly social and cultural constraints than does the appropriate background consensus in science. But these features of moral argument are simply a function of the fact that moral reflection structures experience in ways not coincident with reflection about rabbits, electrons, or black holes.

The structure of moral experience is largely, though not entirely, a function of our self-understandings—particularly, our collective cultural self-understanding. A substantial portion of the background consensus necessary for moral agreement will thus be a function of membership in communities bound by much richer ties than the merely linguistic ties which unite theoretical reasoners. Because such communities will generally be smaller than linguistic communities, the consensus embodied in them will be more narrowly circumscribed than that embodied in linguistic communities. Moreover, my claim that the structure of moral experience is largely a function of collective self-understandings must be properly understood: a moral theory is by no means *about* the content of anybody's self-understanding. A moral theory is a theory about what ought to happen, and, about how we should act, think, perhaps even feel, in order to bring that about. Yet the moral life is possible only as a consequence of self-examination, and the self-conceptions to be subjected to (and possibly revised by) rational scrutiny are the only empirical foothold of ethical theory. Justification in ethics is thus rightly construed as a matter of trying to secure collective agreement on the results of that rational self-scrutiny.

We shouldn't want the methods of ethics to be more like those in science, because the process of theory-formation in science provides no opportunity for self-scrutiny. As Rawls reminds us,

> if we have an accurate account of the motion of the heavenly bodies that we do not find appealing, we cannot alter these motions to conform to a more attractive theory.[33]

But while we cannot alter the motions of the heavens, we can try to see ourselves in a different light. The capacity to treat our self-conceptions as revisable is an essential component of rationality as well as a prerequisite of moral theory. Moreover, the willingness to exercise this capacity is a condition of peaceful co-existence with others, and is the only attitude capable of ensuring that the structure of moral experience fits comfortably with changes in the texture of non-moral experience brought about by social and scientific change. We must therefore resist the effort to consign to the sphere of the non-rational realms of discourse like ethics, which afford opportunities for self-scrutiny. A Quinean may take my account of moral discourse to show that ethics is concerned with data which have no place in a naturalized vision of the world. Yet it is at great cost that we accept that vision as exhaustive of "what there is." Moral theory and moral argument are valuable not merely in spite of, but *because* of, the fact that they rely on methods appropriate to their subject matter—methods fundamentally different from the methods of science.

University of Rochester

Received June 16, 1989

NOTES

1. W. V. O. Quine, "On the Nature of Moral Values," in *Theories and Things* (Cambridge: Belknap Press of Harvard University Press, 1981), p. 63.

2. Owen Flanagan, "Quinean Ethics," *Ethics*, vol. 93 (1983), pp. 56-74.

3. A. J. Ayer, *Language, Truth and Logic* (Middlesex: Pelican Books, 1946; 2nd ed., 1971), esp. pp. 136-58.

4. *Ibid.*, p. 149.

5. C. L. Stevenson, *Ethics and Language* (New Haven: Yale University Press, 1944), esp. pp. 02-08 and 72-73.

6. *Ibid.*, pp. 138-51.

7. Quine, *loc. cit.*

8. *Ibid.*, pp. 64-65.

9. This discussion of Quine's views draws primarily on *Word and Object* (Cambridge: MIT Press, 1960); *Ontological Relativity and Other Essays* (New York: Columbia, 1969); and *Theories and Things* (Cambridge: Harvard University Press, 1981).

10. Quine, "Reply to Morton White," in *The Philosophy of W. V. Quine*, ed. by Lewis Edwin Hahn and Paul Arthur Schilpp (La Salle: Open Court, 1986), p. 664.

11. Quine, "Five Milestones of Empiricism," *Theories and Things*, p. 71.

12. *Ibid.*, pp. 70-71.

13. Quine, "Reply to Morton White," pp. 663-64.

14. It has been argued that there are ethical observation sentences. Gilbert Harmon cites "It is wrong for those kids to torture that cat like that," in his, "Moral Explanations of Natural Facts—Can Moral Claims be Tested Against Moral Reality?" *Southern Journal of Philosophy*, vol. 24 (1986) (Spindel Conference on Moral Realism), pp. 58-59. Flanagan offers a similar example in "Pragmatism, Ethics and Correspondence Truth: Response to Gibson and Quine," *Ethics*, vol. 98 (1988), p. 546. Yet though there is widespread agreement on the theory (or portion of theory) that condemns torturing the innocent, that *theory* is the "collateral information" needed to secure agreement on such sentences.

15. Flanagan, "Quinean Ethics," p. 65.

16. Quine, "On the Nature of Moral Values," pp. 55-57.

17. W. V. Quine and J. S. Ullian, *The Web of Belief* (2nd. ed.; New York: Random House, 1978), pp. 64-82.

18. Quine, "Reply to Morton White," p. 664.

19. *Loc. cit.*

20. David Hume, *A Treatise of Human Nature*, ed. L. A. Selby-Bigge. 2nd ed., revised by P. H. Nidditch (Oxford: Oxford University Press, 1978), Bk. II, Pt. 3, Sec. x, pp. 448-54.

21. Flanagan, "Quinean Ethics," p. 71.

22. Flanagan, "Pragatism, Ethics and Correspondence Truth," pp. 542-43.

23. John Rawls, *A Theory of Justice* (Cambridge: Harvard Univ. Press, 1971), pp. 100-8. One of the "fixed points of our considered judgments," Rawls claims, is "that no one deserves his place in the distribution of native endowments...(p. 104).

24. Immanuel Kant, *Groundwork of the Metaphysic of Morals*, tr. by H. J. Paton (New York: Harper and Row, 1964), pp. 71-72; and Kant, *Critique of Practical Reason*, tr. by Lewis White Beck (Indianapolis, Indiana: Bobbs-Merrill, 1965), p. 8n.

25. R. M. Dworkin, "The Original Position," in *Reading Rawls: Critical Studies of A Theory of Justice*, ed. by Norman Daniels (New York: Basic Books, n.d.), pp. 16-53, esp. pp. 27-37. See also, John Rawls, "Kantian Constructivism in Moral Theory," *The Journal of Philosophy*, vol. 77 (September, 1980), pp. 515-72, esp. 518-19.

26. Gilbert Harman, "Moral Explanation of Natural Facts," p. 61. See also, Daniel Little, "Reflective Equilibrium and Justification," *The Southern Journal of Philosophy*, vol. 22 (1984), p. 379 and p. 386n.

27. Rawls takes a different view of intuitions in, for instance, "The Independence of Moral Theory," *Proceedings of the American Philosophical Association*, vol. 48 (1975), p. 7.

28. Rawls claims to describe our sense of justice in *Theory of Justice*, pp. 46-53. His more recent work—starting with "Kantian Constructivism in Moral Theory"—downplays (or ignores) this notion.

29. Rawls, *A Theory of Justice*, p. 48-49.

30. *Loc. cit.*

31. Consistency and coherence (internally, and with non-moral beliefs) are examples of norms of rational self-scrutiny.

32. I thus depart from Rawls' view in, for example, "Kantian Constructivism in Moral Theory." Rawls writes there that "the 'real task' of justifying a conception of justice is not primarily an epistemological problem," but a "practical social task" (*op. cit.*, pp. 518-19).

33. Rawls, *A Theory of Justice*, p. 49.

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 27, Number 3, July 1990

EVOLUTION, "TYPOLOGY" AND "POPULATION THINKING"

Marjorie Grene

EVOLUTION happened. That's our starting point. How it happened may still be, in part, open to question. But that it happened is a fact, just as it is a fact—debatable in Descartes's and Harvey's time—that the blood circulates. Given organic evolution, then, it is sometimes asked, can there still be a theory, or better, a philosophically or scientifically grounded concept, of human nature, or at least a concept compatible with the state of scientific knowledge? The question is, or ought to be, twofold: first, given organic evolution, can there be "natures" at all? If all things flow, what is a "nature"? Second, can there be *human* nature, which we like to think somehow unique, contrasted to other, "lesser," breeds-without-the-law? Today I want to address only the first half of this question: why, within the framework of evolutionary biology, is there a difficulty about natures in general, and, further, what can we do about it?

It is an article of faith in evolutionary biology that Darwin shifted biological theory and biological research from a static, evil, "metaphysical" kind of thinking called "essentialism" or "typology" to a dynamic, good, scientific conceptual style called "population thinking." This belief, given its most authoritative statement by Ernst Mayr, has usually been alleged to hold peculiarly for the biological sciences in their unique advance to modernity. In a widely acclaimed book Elliott Sober has now generalized the contrast: except for Darwin and the theory of natural selection, all science, he claims, has always been and continues to be held captive by that wicked typological attitude: natural state explanation, as he calls it (Sober, 1984). This seems to me much too sweeping a claim, but perhaps it is worth exaggerating here, if only to stress the alleged uniqueness and alleged explanatory power of "population thinking." What was this one-time, extraordinary shift in the basis of scientific thought?

As every one knows—and forgive me for saying it once more—what Darwin did was to turn the attention of the observer or experimenter to the minute differences between particular organisms, as distinct from the specimen "typical" of the kind, and so to allow us to imagine how, through the gradual accumulation of such tiny variations, new varieties, and finally the larger varieties we call species, might have arisen. That is the first of the three facts of life that define the conditions for natural selection—or perhaps, as some people read the argument, define natural selection itself (but cf. Lloyd, 1988). Of course to produce the dynamic of Darwin's theory, the theory of natural selection, you need (as Lewontin stated in his classic 1970 paper (Lewontin, 1970)) in addition to phenotypic variation two further principles: differential fitness, that is, differential adaptedness to a given environment and heredity, in particular, the inheritance of those differentially fit characters. The second principle raises the large and central issue of the nature and role of adaptedness, and of the relation of "fitness" to adaptedness (Burian, 1983; Brandon, 1981). And the whole scenario leaves open the question of the units of selection (Brandon and Burian, 1984; Wimsatt, 1980; Sober, 1984; Lloyd, 1988). Fortunately, however, we can put aside here those momentous issues, and confine our puzzlement chiefly to the problem, sticky enough, I think, of what it means to concentrate on small variations and dismiss the confining bonds of thinking about *types*.

Typological thinking is in general identified with the Aristotelian tradition, and so it is, of course, in Sober's "natural state" explanations (Sober, 1984). And although there was not necessarily one property, or even a few properties, tagged as essential to a given Aristotelian kind, it was indeed the least or indivisible kind, the *atomon eidos*, for which Aristotle searched, and on the existence and knowledge

of which in his view a proper science had to be based (but cf. Pellegrin, 1985). Much later, in the seventeenth and eighteenth centuries, when the *scala naturae* (a non-Aristotelian notion, be it noted) had become standard, even taxonomists who rejected that perspective and saw nature rather as a complex network of only slightly differing kinds, often spoke of a *degeneration* from an ideal, so that it was some one norm (typically, I fear, the human one) from which all else appeared to deviate (Daudin, 1926). Apart from such cosmologizing taxonomies, moreover, practitioners of natural history from Aristotle through Darwin to present-day observers of birds or butterflies or wild flowers have to learn to tell one *kind* from another, the house finch from the purple finch, the Monarch from the Viceroy butterfly and so on. That's all *typological* practice and underlying it, at first sight at least, there seems to be in all of us—until we learn better—a tendency to typological thinking.

The population thinker, however, *has* learned better. He looks at *this* individual bird or butterfly or daisy and asks, not what archetype it approximates, but how it differs, ever so slightly, from its neighbors. Far from being "careful of the type," as Tennyson pronounced, nature ignores types. Indeed, there *are* no types, only minutely differing individuals aggregating to populations. Populations, if this contrast is to hold, must themselves be indefinitely variegated collections of particulars, each of which differs from each other if only very slightly. What there is is nothing "normal," nothing "typical," but only what Aristotle would call a heap (*swros*), as in a heap of grain. And it is the fact that what is ontologically basic—what really exists—is *heaps* that permits evolution. Every particular just is what it is, from birth to death; but the statistical make-up of aggregates of particulars can change—and that, when it is perpetuated through heredity, is evolution. It is one of the slogans of orthodox evolutionary theory that individuals do not evolve, only populations. (Although Darwin, still susceptible to what we call Lamarckism, was not always sure of this, his descendants are.)

Now if populations are the units of evolution, and populations are simply aggregates of particulars, in no way to be placed into types or judged in accordance with archetypes or essences, it seems to follow that there are no natures. The sharp discontinuities that naturalists observe between many (though of course not all) species are illusory. There is really a perpetual kaleidoscope, almost amounting to a flow, of ever so slightly differing particulars. To talk of *the* nature of *the* lion or *the* grasshopper or *the* ant is to speak typologically, to relapse into the discourse of a murky, pre-Darwinian past.

Or so the story goes. But there are problems. It has often been pointed out, for example, that the general view of variation as eventuating in selection, while it underlies the argument of the book *On the Origin of Species*, does away with the very concept, the very existence, of species. There are only varieties. In short, the origin of species is exactly what the *Origin of Species* is not about (Beatty, 1982; cf. Beatty, 1984, which does not, however, undercut his previous argument as he believes it does!). What concerns me here, however, is not that paradox, but another (perhaps related) one. Population thinking, which focuses on slight differences between particulars, is contrasted with typology, which elevates the norm or kind to first place, and sees particulars insofar as they differ from that norm as deviations from it. Along with the abolition of a deterministic Providence (God's will be done!), that was supposed to be one of the liberations performed by Darwinism. But as we have seen, it is not particulars (individual organisms) that evolve, it is batches of them. If you count fruit flies in a genetic assimilation experiment, for example, it is the relative number of crossveinless mutants that matters, not *this* fly or that one. I could appreciate this individual as a beautiful specimen of the species *baltimore oriole* or the variety *bulldog*; but it is only collections that evolution "cares about." Populations are aggregates to be statistically assessed, not, not ever, individuals. But there is something odd, and oddly unresolved, about a style of thinking that allegedly concentrates on particulars as distinct from types or kinds, only to put its weight on large collections of particulars, populations of birds or butterflies or daisies rather than on *this* slightly differing bird or butterfly or daisy as distinct from that one. Like other traditional atomizing thought styles, "population thinking" wants to build on particulars in order to track changing aggregates of such particulars through time.

That is the starting point for a deeper paradox. When evolutionists talk about slight variations, we may ask, variations of *what*? Of bristle number, or length of limb, or skin pattern or pigmentation, or

what you will, but of some trait or other. Variants must differ from their neighbors, however minutely, in some character or characters. But characters are not and cannot be particulars. They are sortals, predicates that sort out different kinds. Indeed, as Sober himself stresses, it is the (slightly) differing *properties* of organisms that matter in the selection process, in the first instance, the properties *for* which these organisms rather than others are selected, and as a corollary also other, accompanying or consequent characters that happen to be selected (see Sober (1984) on selection for and of, Gould and Vrba (1982) on adaptation *vs.* exaptation; and cf. Waddington 1957 on selection of, for and by). No matter how tiny the variations that selection works on, they must be variations on some theme, variants of some character, which, *in* its variation, makes a difference to the reproductive success of the organism possessing it.

You may say, if you like, that sortals enter here only as necessities of our language or of our thought. But that won't do—since if the characters in the case didn't *really* make a difference to their owners, and didn't make the same difference in the next and next generation, or, if the environment changes, if they didn't make a different difference—in short, if there were not those real effects of differential properties, not only the evolutionist's discourse, but its object, would disappear. "Population thinking" taken strictly turns out to be not only methodologically, but ontologically, impossible. Types, kinds, sorts are bound to crop up somewhere, else we not only could not speak about nature; there would not be a nature to speak about. As Plato put it, it is the intermingling of forms that makes discourse possible for us (*Soph.* 259E). Unless reality contains some order, in other words, we can find no order in it.

There I go again, the evolutionist (I mean an orthodox Darwinian evolutionist, whether a scientist like Mayr or a philosopher like Hull or Sober)—there I go again, the evolutionist would say, thinking typologically like most armchair philosophers before me. Types, these writers hold, or seem to hold, are by their very nature eternal; but the living world is constantly changing. If there is something like "types" in existence, they are illusory, since they have arisen gradually through changing adaptations (or exaptations) from earlier, and different, organisms in earlier and other environments, and they will go extinct to give way to the novel adaptations of new and unforeseen entities in new environments. A "type" or a "nature," however, these thinkers seem to believe, is by its very nature (?) eternal, irrelevant to flux. Insofar as the real world changes, therefore, so far there are no natures, and if it changes altogether, there are no natures at all. A "type," or a "nature," moreover, they proclaim, is defined *intensionally*, with reference to some properties it necessarily possesses (see e.g. Hull, 1976) That's how it shows what it is. But insofar as Nature, the real world of particulars and populations, is evolving, coming to be and passing away, it is impossible to specify any properties whatsoever without which it could not be.

I have always found this argument very difficult to understand. Why, just because something does not last forever, should it lack a nature? People have characters, other animals of a size and in a situation that permit personal acquaintance have characters. Indeed, the protozoologist H. S. Jennings argued that were *Paramecia* as large as dogs (and domesticated?) we would give them names and know them apart, by their traits and dispositions, just as we do our present household pets (Jennings, 1906). That animals are born and die doesn't mean you can't tell one from another while it lasts, or even after it is dead. I remember an exceedingly timorous dog we had long ago in Illinois and another whom we had to get rid of because she was a killer. That was his nature, and that was hers. And why, equally, shouldn't we be able to talk about the nature of a certain collection of plants or animals, better yet, a family, or lineage, of plants or animals? There are no mammoths nowadays, but surely we can say something about the *kind* of ungulates mammoths *were*.

There is another odd angle to this situation, too. In their eagerness to abandon "essentialism," some biologists (and philosophers) have put forth the thesis that species are not kinds at all (which, according to these writers, are tied forever to their essential properties and so could not evolve). Species are individuals (which are simply baptized, tied to no properties, and so can be the subjects of evolution) (e.g. Hull, 1986). But it is not individuals that evolve (that would be Lamarckism, which is just as bad as typology), it's populations. So where are we?

I do not want to go into the species = individuals debate at this point; it is a logical quagmire and I hope I can say what I want to without plunging into that dismal swamp. My point is: that the attempt to

abandon talk of kinds or natures altogether gets us into very strange corners, and we need not after all go that way. On the contrary, an evolving universe can and does throw up things *of* kinds, *with* natures, and we should be able, short of typological pontificating, to say something sometimes about some of them. Of course our acceptance of the fact of evolution has important consequences for our view of the nature of natures, but in itself it does not demand that we abandon altogether the effort to distinguish one life style from another, one kind of plant or animal from another. Evolution gives a new and richer meaning to "nature," it does not abolish it.

How, then, within the conceptual framework demanded by the fact of evolution, are we to interpret "kinds" of living things, or natures? If evolution is not a sheer flux that undercuts altogether any stable "nature" of things, in other words, if evolution stops short of a total destruction of stability, does it alter traditional ways of thinking about the natures of particular kinds of plants and animals? And if so, how?

The answer depends, I believe, on the style of evolutionary theory one accepts. I am thinking here, be it said, only of theories, or, more vaguely, perspectives, that belong in the most general sense to the Darwinian tradition.

Perhaps a schematic division of evolutionary theories will help us orient ourselves here. Such theories tend to explain the causality of evolutionary change either *endoetiologically*, referring to some aspect of organisms themselves as instrumental in bringing about the extinction of some species and the origin of others, or *exoetiologically*, with reference to environmental change as the chief motor of evolution. Before the nineteen thirties endoetiological theories were common: like Berg's "nomogenesis" or Osborn's "aristogenesis," views that attributed to organisms themselves some inner drive toward the development of new and "higher" forms. Nowadays some novel internalist views are once more being touted, this time however, theories that attempt to assimilate biological change to general physical laws with implications for the origin and development of systems in general, including living systems. Over against either of these versions of endoetiological explanation, there have been two major forms of exoetiological theory: in our century, neo-Lamarckism and neo-Darwinism (or the synthetic theory). Neo-Lamarckism claimed to have found a major cause of evolution in the direct influence of the environment and of the organism's response to it, accepting the inheritance of acquired characters. Neo-Darwinism is equally externalistic in its basic explanatory thrust, finding the motor of evolution in the relation of organisms to a given environment and rejecting firmly any kind of internal drive to evolutionary change, orthogenetic, aristogenetic or what you will. With the advance of genetics and the increasing influence of Weismann's principle: the one-way boundary between genotype and phenotype, neo-Lamarckism died away and what was left was a two-step, Darwinian, theory. There is variation which is random with respect to the needs of the organism, and in a given environment there is, in view of such variation, a differential adaptedness of individuals within a population and hence differential reproduction, leading to change in the statistical make-up of the population, hence evolution. That is the core of contemporary Darwinian thinking (Wright, 1967). And it is within that very flexible framework that I want to consider alternative views of "nature" or "natures." In other words, I am ignoring recent revivals of endoetiology, the chief articulate alternatives at the moment to the Darwinian version of exoetiological theory. (I am also omitting consideration of the theory of "neutral mutations," with apologies to any of its adherents who believe it constitutes a viable alternative theory.)

Let us ask, then, how, within the Darwinian tradition, broadly understood, we can describe or analyze the "nature" of some particular kind of organism. Now within this generic thought-style (as Fleck would call it; Fleck, 1981), there are variants of emphasis that make a difference for our problem. Versions of Darwinian theory can be more or less linear, less or more multi-dimensional in their concepts and principles and therefore in the consequences they entail.

1) The most linear is genic selectionism, the view that a) evolution *is* natural selection and b) natural selection is to be identified with changes in gene frequencies. We have to look at the fate of phenotypes—individual organisms—because in nature, with our limited perceptions, we cannot see all the way down to the genes; but it is genes copying themselves, more or less successfully, that evolution *is*. As Samuel Butler put it long ago, in a tag that Dobzhansky quoted approvingly, a hen is just an egg's way of making another egg. G. C. Williams

has given the classic modern statement of this view (Williams, 1966); its most extreme recent version can be found in such works as Richard Dawkins' *Extended Phenotype* (Dawkins, 1982). The concept of the phenotype is here taken very broadly, yet its significance is still purely instrumental: it is a device for perpetuating genes. Evolutionary biology will, or ought to be, transformed into molecular biology, or better yet molecular biophysics. On such a view "evolutionary natures" are a piece of nonsense—like incorporeal substances for Hobbes. Our question cannot meaningfully be asked, let alone answered.

2) Less atomistic, but still reducing evolutionary to genetic processes was Mayr's 1963 definition: "natural selection is simply the differential survival of genotypes," where "genotype" was taken as equivalent to genome, that is, the sum total of an individual's genetic endowment (Mayr, 1963). Mayr insists on the unity of the genotype, so that it is, though at the genetic level, a complex whole that is differentially perpetuated. Although, as in most evolutionary theories so far, development is a black box—one just takes it that genotype A produces a corresponding phenotype A_1—this view does allow a concept of organization into its evolutionary scenario, as extreme genic reductionism does not. But it is still genes, not organisms, let alone organisms whose natures we could describe or think about, that are selected.

3) At the same time, Mayr insisted that it is phenotypes, individual organisms, that are the *target* of selection (Mayr, 1963). This is a point that was also made recurrently and emphatically by C. H. Waddington (e.g. Waddington, 1957). After all, it is the success or failure of phenotypes in reproduction that determines which genes, or genotypes, will be perpetuated. We need at least two levels, organisms and genotypes, or better three, organisms, genotypes and genes, in our account of the evolutionary process. As Wimsatt has put it, the genes do the bookkeeping for evolution; but they are not its executives or even its workers on the factory floor (Wimsatt, 1980).

Long before the establishment of population genetics, that is how Darwin thought about natural selection and *a fortiori* about evolution. (Not that he considered natural selection the only process constituting descent with modification; but it was certainly a major, and perhaps the major, factor.) How do "natures" fare on this classic Darwinian model?

They fare a bit better than with variants 1) and 2) but still not well enough for our question to be posed, let alone answered. At least we can now deal with organisms rather than gene pools or genotypes, but we are still in the ball park of "population thinking"; so we have no right, on principle, to say that *the* lyrebird is "perhaps the most gifted vocalist in the world" or that *the* female huia bird had a slender curved beak twice as long as that of the male. On radical "Darwinian" grounds, however, there is *one* kind we *are* allowed to talk about, and that is, the kind constituted by an ancestor-descendant relation. On this view, organisms themselves are of course traitless, not characterized by "essences" of any sort—except for this one: they do have the property of being united in ancestor-descendant relations. So we can talk about genealogies or lineages, collections constituted by the property of being so related. Granted, except for those of us who remember giving birth, or perhaps for paramecium watchers, ancestor-descendant relations are highly abstract properties. We take them on faith, just the faith Darwinism permits us, whether at the genic, genotypic or organismic level. On this ground, if we take it seriously, "natures" (except for the nature of genealogies!) relapse into "typology." We are inclined to dismiss the very thought of such essentialist nonsense, and to concentrate on the pure linear question, who gave birth to whom? The "who's" involved, however, are mere pointers, pinpoints on a complicated but always two-dimensional diagram. Like the resulting genealogy itself, they have no meaning. Talk of natures is still taboo.

4) How can we go further? Surely the most fundamental lesson of evolution is that natures are histories, that they have origins and endings, but not that they do not exist. Here, fortunately, recent expansions or perhaps modifications of the Darwinian tradition can come to our aid. In order to exhibit these (partly) new developments, unfortunately, I have to supplement my story so far in two different directions. I say "partly" new, because one of them certainly can be found in Darwin, while the other is, I believe, genuinely innovative. The trouble is that the two interact with one another, so that my separation of dimensions in this multi-dimensional, expanded Darwinism results from the linearity of discourse, not of reality. These two complexifying factors are: first, taking ecology seriously *within*

evolutionary theory and second, introducing a hierarchical perspective into evolutionary biology.

As I have already indicated, Darwinian explanation is exoetiological. The moving force of evolution comes from the adaptive relations of organisms to their *environments*. In a stable environment, such as population-genetical models classically deal with, it is differential adaptedness to that environment that triggers evolutionary change. In a changing environment evolution results from the slightly better adaptedness of some organisms than others to new conditions. An ecological approach is fundamental to Darwinian explanation. Yet the science of ecology has developed in this century largely independently of—and sometimes even in seeming opposition to—Darwinian modelling based on population genetics (Kimler, 1983). By now it is clear to many that there is more than just genealogy, whether gene sequences or even a phenotypic sequence of parent-offspring relations, involved in the evolutionary process. A classic statement supporting this insight, for example, is Hull's 1980 paper, in which he distinguished "replicators" and "interactors," arguing that it is through the interactions between these two that lineages, and hence evolution, are produced (Hull, 1980). Strictly speaking, only genes replicate, but if, following Eldredge, we generalize "replicators" to "more-makers," and "interactors" (or better, "interactions") to "matter-energy exchange," we can see that individual organisms in fact spend their lives engaging in activities of both sorts, reproduction on the one hand and economic life on the other (Eldredge, 1986; Eldredge and Salthe, 1984). Granted, the males of some arachnid species are little but machines for passing on genetic information, and sterile female workers in many insect species keep the economic life of the community running without any (direct) input into genealogy. But in general it is important to an adequate view of the evolutionary process to acknowledge both these tasks: to distinguish information transfer from economics, or matter-energy exchange. Although either of these activities can be studied relatively synchronically or diachronically, we can perhaps think at least analogically of information transfer as running through time, from one generation to the next (or, indeed, from one series of generations to another such series), while the economic life of populations of a given species takes place in a cross-section of history through interactions of conspecifics or between organisms from different species or between organisms and their abiotic environments. Of course reproduction needs space and economic life takes time, but the temporo-spatial distinction may perhaps afford us a crutch to help understand this crucial distinction. Information transfer occurs from one generation to the next; matter-energy transfer occurs between contemporaries, whether of the same species or different species (or between living things and their physical environment).

This sorting out of the genealogical from the ecological, giving each its due weight, consists in a way (as I have already suggested) in drawing out more clearly a theme already implicit in the *Origin*. Try analyzing out, for example, the various inputs and the various effects involved in Darwin's example of cats, mice, bees and clover. The distinction between information transfer and matter-energy transfer, however, as I have sketched it here, is featured in some versions of a more general innovation in recent biological thought, one that is less clearly "Darwinian"—and that is the development of what is sometimes called hierarchy theory. The term "hierarchy" has been used in a confusing variety of ways in recent evolutionary biology and systematics; I cannot stop to sort these out here (Grene, 1987, 1988). Staying with the information-economics distinction, let us just see how hierarchical thinking enters into each.

On the genealogical side, the easiest entrèe is through the controversy about units of selection. The target of selection, we have seen, may be understood to be the gene, the genome, or the organism. Yet even Williams, the arch-defender of genic selection, admits that sometimes, as in the *t*-allele in the house mouse, whole demes (i.e., small groups of conspecific neighbors) are selected (Williams, 1966; Lewontin and Dunn, 1960). And sometimes one wants to distinguish also between species that speciate more or less rapidly, taking the species itself, if not as the unit of selection, at least as the unit of a sorting process through evolutionary time. (A "species" here is taken to be a long chain of information transfers between one generation and the next of potential interbreeders.) Thus species, too, it seems, make more of themselves. And so, if you like, does any monophyletic taxon. In short, there is a genealogical hierarchy, consisting of a series of nested entities from genes to monophyletic taxa.

On the ecological (interactive) side, moreover, the

hierarchical array is equally clear. Organisms exist in populations of conspecifics, which form communities with neighbors of other species: predators, prey, competitors for scarce resources, and so on. These in turn are contained in ecosystems, which sum up to the whole biota organized ultimately in conjunction (and interaction) with the inorganic as well as the living environment. Eldredge and Salthe have schematized these two hierarchies as follows:

GENEALOGICAL HIERARCHY	ECOLOGICAL HIERARCHY
Codons	Enzymes
Genes	Cells
Organisms	Organisms
Demes	Populations
Species	Local ecosystems
Monophyletic taxa	Biotic regions
(Special case: all life)	Entire biosphere

(Eldredge and Salthe, 1984; cf. Eldredge, 1986; Vrba and Eldredge, 1984. We may ignore here differences in various formulations of the two hierarchies; it is the separation of the two that matters for the present purpose.)

Evolution, finally, once we have made these distinctions—between information-transfer and matter-energy exchange as well as between nested levels of each—evolution is seen to result from the interaction between the two hierarchies. Species provide the units—organisms—for the economic arena, and the economic life of organisms determines which units will indeed pass on information to the next generation, thus altering the character of the participants in the economic game at the next round, and so on.

Although the sorting out of these aspects of the evolutionary process has barely begun, I hope we can see that (at long last!) we have opened a way here for the consideration of the *nature* of a given (kind of) organism within an evolutionary perspective. Ordinarily, because of our particular size and life span, we like to look at, and think about, individual organisms of certain species. That means individuals that form part of a given ancestor-descendant network and are characterized within any given generation by a describably and intelligible life-style. Within limits, of course, such a life-style develops, but between origin and extinction of the species it remains recognizably "the same." Thus evolutionary biologists studying a particular species of, say, lizards in the Mojave desert, or of harvester ants in Arizona, can describe and analyze the "typical" activities of metabolism, growth, locomotion, temperature regulation (matter-energy transfer) characteristic of the individuals that instantiate this species in a given year or season or month or day. This is not in any vicious sense "typology"; indeed, it is the study of the bricks and mortar of which evolution is built. True, modern ecologists don't talk about "natures" as the writers of medieval bestiaries did; nevertheless when they study the behaviors of organisms in their environments and the interacting environmental dynamics that influence and even generate such behaviors, it is life-styles, that is, natures, natural ways of being, that they are investigating. Thus a hierarchically enriched evolutionary ecology provides, I would suggest, a much richer perspective for studying the characters of members of a given species *qua* members of that species, and as a product of evolution, than was available in earlier centuries, not only to writers of bestiaries, but even to great naturalists like Louis Agassiz or John Ray, who lacked the evolutionary perspective that we, fortunately, have at our disposal.

Virginia Polytechnic Institute and State University

Received July 13, 1989

NOTES

*This paper was presented at a conference in honor of the late Joan Kung at the University of Wisconsin, Madison in February, 1988 and I would like to dedicate it to her memory. I am grateful to Dr. David Winkler for his careful criticism of my manuscript.

REFERENCES

Beatty, John (1982) "What's in a Word? Coming to Terms in the Darwinian Revolution." *Journal of the History of Biology*, vol. 15, pp. 215-39.

Beatty, John (1984) "Speaking of Species: Darwin's Strategy," in D. Kohn, (ed.), *The Darwinian Heritage* (Princeton: Princeton University Press, 1984), pp. 265-81.

Brandon, R. N. (1981) "A Structural Description of Evolutionary Theory," *PSA 1980*, vol. 2, pp. 427-39.

Brandon, R. N. and R. M. Burian, (eds.) (1984) *Genes, Organisms, Populations* (Cambridge, MA: Bradford Books, 1984).

Burian, R. M. (1983) "Adaptation" in M. Grene (ed.), *Dimensions of Darwinism* (Cambridge: Cambridge University Press, 1983), pp. 286-314.

Daudin, H. (1982) *De Linné à Jussieu* (Paris: Alcan, 1926).

Dawkins, Richard (1982) *The Extended Phenotype* (San Francisco: Freeman, 1982).

Eldredge, Niles (1986) "Information, Economics and Evolution," *Annual Review of Ecology Systematics*, vol. 17, pp. 351-69.

Eldredge, N. and S. N. Salthe (1984) "Hierarchy and Evolution," *Oxford Surveys of Evolutionary Biology*, vol. 1, pp. 182-206.

Fleck, Ludwik (1981) "On the Question of the Foundation of Medical Knowledge," (tr. T. J. Trenn), *Journal of Medicine and Philosophy*, vol. 6, pp. 237-56.

Gould, S. J. and E. Vrba (1982) "Exaptation—A Missing Term in the Science of Form," *Paleobiology*, vol. 8, pp. 04-15.

Grene, M. (1987) "Hierarchies in Biology," *American Scientist*, vol. 75, pp. 504-09.

Grene, M. (1988) "Hierarchies and Behavior," in *Evolution of Social Behavior and Integrative Levels*, proc. 3rd T. N. Schneirla conference (Hillside, NJ: Erlbaum, 1988), pp. 3-17.

Hull, David (1976) "Are Species Really Individuals?," *Systematic Zoology*, vol. 25, pp. 174-91.

Hull, David (1980) "Individuality and Selection," *Annual Review of Ecology and Systematics*, vol. 11, pp. 311-22.

Jennings, H. S. (1906) *The Behavior of the Lower Organisms* (New York: Columbia University Press, 1906).

Kimler, W. C. (1983) "Mimicry: View of Naturalists and Ecologists Before the Modern Synthesis," M. Grene (ed.), *Dimensions of Darwinism* (Cambridge: Cambridge University Press, 1983), pp. 99-127.

Lewontin, R. C. (1970) "The Units of Selection," *Annual Review of Ecology and Systematics*, vol. 1 (1970), pp. 1-18.

Lewontin, R. C. and Dunn, L. C. (1960) "The Evolutionary Dynamics of a Polymorphism in the House Mouse," *Genetics*, vol. 45, pp. 705-22.

Lloyd, E. A. (1988) "Evaluation of Evidence in Units of Selection Controversies," *Philosophy of Science* (in press).

Mayr, Ernst (1963) *Populations, Species and Evolution* (Cambridge, MA: Harvard University Press, 1963).

Pellegrin, Pierre (1985) "Aristotle: A Zoology Without Species," in A. Gotthelf (ed.), *Aristotle on Nature and Living Things* (Pittsburgh: Mathesis Pub., 1985), pp. 95-116.

Sober, Elliott (1984) *The Nature of Selection* (Cambridge, MA: Bradford Books, 1984).

Vrba, E. S. and Eldredge, N. (1984) Individuals, Hierarchies and Processes: Toward a More Complete Evolutionary Theory. *Paleobiology*, vol. 10, pp. 146-71.

Waddington, C. H. (1957) *The Strategy of the Genes* (London: George Allen and Unwin, 1957).

Williams, G. C. (1966) *Adaptation and Natural Selection* (Princeton: Princeton University Press, 1966).

Wimsatt, W. C. (1980) "Reductionistic Research Strategies and Their Biases in the Units of Selection Controversy," in T. Nickles, ed., *Scientific Discovery: Case Studies* (Dordrecht: Reidel, 1980), pp. 213-59.

Wright, Sewall (1967) "Comments on the Preliminary Working Papers of Eden and Waddington," in *Mathematical Challenges to the neo-Darwinian Interpretation of Evolution*, Wistar Symposium Monograph, no. 5, pp. 117-20.

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 27, Number 3, July 1990

EPISTEMIC INTERNALISM'S DILEMMA

Stephen Cade Hetherington

PART 1. THE DILEMMA

I

SUPPOSE some epistemic internalist believes that you have a justified belief.[1] At least part of what he or she thinks is that *some* aspect *A* of your circumstances is epistemically internal to you and is at least *part* of what makes you justified.[2] In thinking that *A* is epistemically internal to you, the epistemic internalist is presumably crediting you with some *grasp* of *A*.[3]

But what does this "grasp" involve? For a start, the internalist *should* not just be assuming that *A* is *mentally* internal to you (in the sense that if you *have* a belief it is mentally internal to you). Being mentally internal, say, should not be seen as sufficient for being epistemically internal.[4] The same is true of your being aware *that* the item is mentally internal (e.g. your being aware that you have the belief in question): even this should not be viewed as sufficient to make the item epistemically internal. For example, let *A* be a belief of yours that an epistemologist could think helps justify some other belief *B* of yours, and suppose you grasp *A*, in that you grasp its *presence*. But suppose, too, that you have no idea that *A* plays a role in making *B* justified. I suggest that, in such a case, *A* is no more *epistemically* internal to the *epistemic* situation constituted by *A* helping to justify *B* than is any *other* belief *C* you realise you have—where an epistemologist looking on might see *C* as *irrelevant* to *B*'s being justified. *C* is just as mentally internal to you, say, as is *A*; but this should hardly be sufficient for its being epistemically internal. If I am right, then, the epistemic internalist should conclude that *A* is *not* epistemically internal to you (even though it is mentally internal to you). For then the following structure is instantiated:

(E) Something is helping you have justification, even though you have no idea, awareness, or appreciation *that* it is doing so.

Instantiating (E) should be viewed as what *makes* the contemporary paradigms of epistemic externalism epistemically externalist. Thus, someone like Alvin Goldman (1979) would insist that part of what makes your belief justified can be its being held as the result of a *reliable* belief-forming process—even when you have no idea *that* this is true of your belief. (E) is thereby instantiated; an epistemically *external*ist condition of justified belief is present.[5]

Equally, what makes Descartes the traditional paradigm epistemic internalist should be seen as the fact that his view of something's contributing to your certainty does *not* instantiate (E). He asks not only that your idea *be* clear and distinct, for instance; he wants you to appreciate *that* it is. And in his fourth *Meditation* he asks that you be continually *aware* of the principle that a clear and distinct idea is true; the *truth* of this principle is not enough. God can supposedly make a clear and distinct idea be true, but God does not determine what judgements Descartes has. Descartes's search for true judgements involves his appreciating the role of clarity and distinctness in attaining that end. It also requires that he not simply rely on the *truth* of the belief that God exists and is not a deceiver; he thinks he needs to be aware *that* God exists and is not deceiving him.

II

Central to epistemic internalism, therefore, is the view that *A* is epistemically internal to you only if you can appreciate it as such. Even if *A* is internal to you in, say, a mental sense, it still might not be *epistemically* internal to you. The epistemic *internalist* might purport to describe your situation, from his or her epistemological perspective, as including your epistemically internalising *A*—but unless *you* appreciate *A* as helping you have justification, *A* will

not be epistemically internal to you. (It depends on you; it is *your* epistemic Within, so to speak.) If *A* can contribute in way *W* to your being justified,[6] without your being aware *that* it is doing so, then *A* instantiates (E): it is, like the fact of your belief being—or not being—reliably formed, epistemically external to you.

A necessary condition, therefore, of some given *A* being epistemically internal to you is your appreciating *that* it is contributing to your being justified. But what about this appreciating, this being aware, of *A*'s contributing? If it is necessary to *A*'s-being-epistemically-internal-to-your-being-justified, then it plays some role, too, in your being justified. (For example, if *A*'s-being-epistemically-internal-to-your-being-justified is necessary to your being justified, then, by the transitivity of necessary conditions, your appreciating of *A* as epistemically internal is itself necessary to your being justified.) Then, however, we must again ask whether (E) applies to the situation. Your appreciating-that-A-is-contributing-to-your-being-justified is contributing to your being justified; need you in turn appreciate *that* it is doing so? What the epistemic internalist should say is that you must. Otherwise, by its instantiating (E), your appreciating-that-*A*-is-contributing-to-your-being-justified is itself epistemically *ex*ternal to your being justified. Yet a necessary condition of *A*'s being epistemically internal to you was your appreciating *that* it is contributing to your being justified. And if the latter aspect of your situation is epistemically *ex*ternal to your being justified, then no doubt the former is too.[7]

The epistemic internalist therefore must confront a dilemma, as follows.

The Dilemma's First Horn. If *A* is to be epistemically internal to you, then you must appreciate *that* *A* is contributing to your being justified. But then you must appreciate that this appreciating is itself contributing to your being justified. And hence, by analogous reasoning, the same is true of the *new* appreciating. This pattern recurs, and the epistemic internalist therefore faces the prospect of an infinite regress of appreciatings, one I assume to be vicious.[8] The regress, (R), is this:

> According to epistemic internalism about a condition *A* of your having a justified belief,
>
> (1) *A* contributes to your justification
>
> only if
>
> (2) You appreciate that *A* contributes to your justification
>
> only if
>
> (3) You appreciate that your-appreciating-that-A-contributes-to-your-justification (i.e. the appreciating which is (2)) contributes to your justification
>
> only if
>
> (4) You appreciate that the appreciating which is (3) contributes to your justification
>
> only if
>
> (5) You appreciate that the appreciating which is (4) contributes to your justification
>
> only if... And so on.

Note that this does not entail that *you* appreciate the "only if" connections; arguably, epistemic internalists themselves must appreciate these, but you need not. Hence you need not appreciate *that* there is an infinite regress either. Nevertheless, I assume that the epistemic internalist's thinking that (R) will be *true* of you (even if you are unaware that it is) will make *him or her* want to avoid (R). He or she should therefore wish to evade The Dilemma's First Horn too.

The Dilemma's Second Horn. On the other hand, if any member of (R) is allowed to be epistemically *ex*ternal to you, then *A* is epistemically external to you also. If, say, (2) is false of you, then so is (1), according, at any rate, to epistemic internalism. Hence the epistemic internalist will decide that *A* does *not* contribute internalistically to your justification. The epistemic internalist will therefore be conceding the fight over *A*'s epistemic role to the epistemic externalist. A necessary condition of an epistemic internalism about *A*'s role in your having justified belief is therefore that each of (1), (2), (3), etc., obtain of you. That is, epistemic internalism returns to The Dilemma's First Horn. But The First Horn *also* implies that *A* will not be epistemically internal to you. On either The First Horn *or* The Second Horn, then, *A* fails to be epistemically internal to you as a justified believer. I call this Epistemic Internalism's Dilemma.

III

The Dilemma says, in effect, that once you *begin* epistemically internalising, you are logically com-

mitted to never *stopping*. The Dilemma is that, since *A* is epistemically internal to you only if you appreciate *that* it is, this commits you to a regress of iterated appreciatings, *unless* one of these appreciatings is epistemically *ex*ternal to you. The only way this pattern of appreciating can avoid regress is if some part of it also avoids being what it was originally supposed to be—namely, part of what made *A* epistemically internal to you. Epistemic internalism must give way to epistemic externalism, and therefore it must be an empty concept. Necessarily, there are no epistemically internal conditions of justified belief.

The Dilemma conjures up an interesting picture of the epistemically internal. It suggests that *A* is epistemically internal only if your appreciating *A as* epistemically internal entails that *A* is therefore *not* epistemically internal. Its being epistemically internal requires your appreciating it as such, but the very appreciating of it as such *destroys* it as such. It might not be destroyed in every sense (it might continue to be mentally within you), but it will cease to play an independent *epistemic* role. *In the very act of being appreciated as epistemically internal*, the purportedly epistemically internal *A* dies, becoming the appreciating-of-*A*-as-epistemically-internal.[9] A *non*-epistemically internal *A* might remain (again, for instance, you might still have a belief *A*), but it cannot be simultaneously internal and *epistemic*.

The concept of an epistemically internal condition of justified belief is therefore like the concept of a reachable horizon. It is necessarily empty. Its being *non*-empty requires that some aspect *A* of your situation be subjected to irreconcilable demands. For the concept of something being epistemically internal to you is *non*-empty only if you focus your thought on *A* in a way which makes that concept *empty*. (Now you see it; therefore you do not! Now you are at the horizon; therefore you are not! You can only reach what *was* the horizon, never what *is* the horizon.)

Part 2. Some Historical Intimations of the Dilemma

I

Part 1 develops the Dilemma in full generality. *A* was *any* aspect of your circumstances which *some* internalistically inclined epistemologist might claim plays *some* role in your having a justified belief. Part 1 also, I think, reveals how *simple* the Dilemma is. It is therefore rather puzzling why epistemic internalism seems so attractive to so many good epistemologists (e.g. BonJour 1985; Chisholm 1989, Ch. 8). This puzzlement should be reinforced by the realisation that (i) views which discuss what a contemporary epistemic internalist could plausibly endorse as possible instances of *A* have a significant philosophical pedigree, and that (ii) *worries* about those possible instances of the epistemic internalist's *A*—worries which do not require much modification in order to be seen as instances of the Dilemma—have an equally significant philosophical lineage.

I will discuss four examples which exemplify (i) and (ii), all from earlier this century. In Wittgenstein's attack on private language, Russell's views on the epistemic accessibility of logical particulars, Sellars's critique of the Given, and a worry raised by Ayer about basic statements, I see more specific voicings of objections which, when adapted to the concept of the epistemically internal, instantiate the Dilemma.[10]

II

Let us consider Wittgenstein's criticism in his *Philosophical Investigations* (1958) of the idea of someone attempting to use a private language. At present there is a great deal of debate about just how to interpret Wittgenstein's argument, and I will not attempt to settle that matter here.[11] I will quickly sketch, instead, how Wittgenstein's argument might be made to play a role in an argument against the possibility of your being able to epistemically internalise a sensation. That latter argument will be seen to instantiate the Dilemma.

Suppose you are a putative private language user: you claim that you can use language to refer to some internal aspect of you (such as a sensation *S*), about which no one *else* can know (*ibid.*, 243). Now, such use is certainly not sufficient for your epistemically internalising *S*: your using a private name to refer to a private sensation *S* does not imply your being aware of *S*'s role in your having a justified belief. So *S*'s being privately referred to is not identical with *S*'s being epistemically internal, and Wittgenstein's private language argument is not automatically an argument against epistemic internalism. However, we can ask whether, when you *are* aware of *S* playing a role in your having a justified belief (as epistemic internalism might require of you), your

awareness includes your use of a private name for *S*. *If* it does, then the tension Wittgenstein notices would be part of the tension the Dilemma captures, as I will now show.

On one interpretation of Wittgenstein (Kripke 1982), the objection to your ability to use a private name for your sensation is that this must presuppose the existence of other *persons* sharing your language, in which case your use of the language is not private after all. On another interpretation (McGinn 1984, 89-90), the objection is that you will have to presuppose the use of the language at other *times*, in which case your present use of the language is not private after all. (So, on the former interpretation, privacy is privacy from other *people*, while, on the latter interpretation, privacy is *temporal* momentariness.) Either way, the structure of the worry portrays the concept of a private language to be flawed much as that of a reachable horizon is. For (says the worry) the attempt to refer to what is private is impossible: a necessary condition of referring to it is that, insofar *as* it is referred to, it is *not* private. And therefore, *if S*'s being epistemically internal to you requires that you have a private, linguistic, grasp of *S*, then part of your instantiating the Dilemma is your instantiating a Wittgensteinian kind of objection.[12] Once again, nothing can be internal and epistemic.

III

One of Bertrand Russell's interests at one time was, of course, the notion of a logical particular (Russell 1918, 55-60). A logical particular is a very short-lived sense-datum, which can be an immediate object of awareness (or, in Russell's terminology, acquaintance). Russell's own motivation for introducing the concept is part of a quest more for a logically perfect language than for an account of justified belief. Still, just as part of one way to possibly instantiate the Dilemma is to consider a Wittgensteinian worry about your epistemically internalising a private sensation by, in part, privately naming it, I will briefly show why I think Russell would endorse the application of Epistemic Internalism's Dilemma to logical particulars. Suppose you think that you can epistemically internalise a Russellian logical particular. Russell says something which suggests that, on grounds similar to mine, he would think you could *not* do so.

If you can epistemically internalise a logical particular, it is only by mentally stepping back from it, appreciating *that* it is internal to you and *that* it can play a justificatory role for you. But when Russell talks of your being acquainted with a logical particular, he clearly precludes your knowing (as he puts it; but we could speak, more generally, of your being aware of) any propositions *about* the particular (Pears 1981, 153). Interacting with complete propositions in this way would bar your interacting with just the logical particulars within those propositions. Epistemically internalising a logical particular would be a *more* complex understanding than the Russellian acquaintance with the particular by itself would be.

Why would this greater complexity "bar" your interacting just with the logical particular, though? You *can* interact with the particular, after all; you can be acquainted with it. Still (from the end of the previous paragraph), this acquaintance will not include your being aware of its *epistemic* role. For *that* awareness could only be *of* a proposition, not a particular. Even if the particular is part of the proposition, the *acquaintance* with the particular itself would not be part of the *understanding* of its justificatory role which would be what the epistemic internalist sees as justificatorily relevant. The acquaintance with the particular would drop out of the picture as epistemically irrelevant. The more complex understanding would take its place, as far as the epistemic internalist is concerned. For epistemic purposes, therefore, the understanding of the particular would supplant the particular. The logical particular can remain internal to you, but it could not be *epistemically* internal to you.[13]

IV

At the beginning of his well-known critique (1963) of the classic foundationalist view of sense-data as the foundations of knowledge (as epistemic givens), Wilfrid Sellars apparently echoes Russell's thinking.[14] If you have a sense-datum, then you sense a particular (*ibid.*, 128). If this is an understanding, or a knowing, of the particular, it is a kind of knowledge that is at least analogous to being *acquainted* with the particular, to use the Russellian term (*ibid.*, 130). However, the epistemic *foundationalist's* purpose in using the sensory given is to ground all propositional empirical knowledge—all empirical knowledge *that* such and such is the case, all knowledge of empirical *facts* (*ibid.*, 128). And that can be done only if the grounds are

themselves cases of propositional knowledge (*ibid.*, 128-129, 131-132). Hence, it is only empirical knowledge of facts, not of particulars, that could be epistemically basic. And, says Sellars (*ibid.*, 134), this is why it is a confusion to think "that a sensation of a red triangle is the very paradigm of empirical knowledge."

And this thinking, when it is applied to the concept of the epistemically internal, instantiates the Dilemma. Let the putatively foundational grounds of empirical knowledge be conscious *awarenesses* that the sensation is present and epistemically relevant; that is, let them *not* instantiate (E), and therefore let them be epistemically internal to you. The problem is that, when the sensation is grasped by you in a way that could be appropriate for grounding your empirical knowledge, it ceases to be relevant *as* a particular. At *that* moment, all that is epistemically operative is the *propositional* awareness. The sensation might be *in* you, but it will not be an *epistemic* part of you. It might be internal, but it could not be *epistemically* internal.

V

In what is presumably his best-known book (1946), A. J. Ayer also employed such thinking. His early positivist's quest for the verifiable led him to what he called *basic* propositions, ones which "refer solely to the content of a single experience" (*ibid.*, 13). And these, he notes, while possessing the apparent virtue of being irrefutable, forfeit this advantage by seemingly committing the sin of not *saying* anything either (*ibid.*, 120-121):

> [These ostensive propositions] are supposed to be purely demonstrative in character, and so incapable of being refuted by any subsequent experience. ... [But] the notion of an ostensive proposition appears to involve a contradiction in terms. It implies that there could be a sentence which consisted of purely demonstrative symbols and was at the same time intelligible. And this is not even a logical possibility. ... The fact is that one cannot in language point to an object without describing it.

Ayer's Introduction (*ibid.*, 14) to the second edition of *Language, Truth and Logic* is less impressed by this first edition reasoning. Nevertheless, it seems right to me, at least when we ask (much as we did for Wittgenstein's private language user) whether an epistemic internalist could coherently conceptualise you as using basic statements to linguistically articulate something's being epistemically internal to you.

Thus, consider these remarks of Ayer's (*ibid.*):

> the form of words that is used to express a basic proposition may be understood to express something that is informative both to another person and to oneself, but when it is so understood it no longer expresses a basic proposition.

These comments suggest that Ayer would agree with the following. If a basic proposition of yours is to refer to some basic experience that is *epistemically* internal to you right now, then you must be able to appreciate, right now, *that* the experience is internal to you and *that* it is contributing to your being justified. However, to have the latter appreciation right now is to supplement the basic experience in a way that effectively *eliminates* it from the epistemic state of affairs. For your experience right now is therefore partly constituted by your mentally stepping back from the supposedly basic experience and using, not the mere *fact* of the basic experience, but your *awareness* of it. The *epistemic* internalising of a basic proposition implies that the basic proposition itself plays no epistemically internalist role.[15]

West Virginia University

Received June 26, 1989

NOTES

1. Epistemic internalists usually discuss justification, not knowledge. The main difference between these two notions is, of course, typically taken to be that knowledge does, whereas justification does not, imply truth. And truth is generally assumed to be an epistemically external factor (e.g. Audi 1988, 113-116). For an overview of the epistemological community's treatment of the epistemic internalism/externalism distinction, see Alston 1986. (Cf. Fumerton 1988 for a slightly non-standard, more recent, view of epistemic internalism.)

2. I say "at least *part*" because, for generality's sake, I am happy to allow that an epistemically internalist condition could be necessary, sufficient, or even something looser still. I am interested in whether something could be epistemically internal

by being *at all* relevant to your having justified belief in some particular context.

3. By "grasp" (or by "appreciate" or "be aware," terms I will use interchangeably) I mean whatever conscious mental operation the epistemic internalist requires you to undergo, in order for you *to* epistemically internalise *A*. This operation might be more—it might be less—cognitively sophisticated. (BonJour (1987) would also talk of *grasping A* (*ibid.*, 297), as well as of having an "inkling" of *A* (*ibid.*, 303) and of whether *A* is your *reason* (*ibid.*, 297, 303, 312n2).)

4. Conversely, it should not be taken to be necessary either: *A*'s being epistemically internal should not be seen as entailing *A*'s being mentally internal. When I say that *A* is some aspect of your circumstances, I *mean* this—in all its apparent generality. Your *grasp* of *A* will presumably be mentally internal (see n. 3 above), but *A* itself need not be. Most epistemologists would prefer that it be mental, of course; the examples I will discuss in Part 2 are all like that. I am only noting that *A* itself *need* not be. I suspect that it is by forgetting this—and so not properly separating the epistemically internal and, for instance, the mentally internal—that epistemic internalists have given us a view which falls prey to my Dilemma. (I suspect, too, that when an epistemic internalist's view *apparently* avoids my Dilemma, it is because the theorist has not properly distinguished what makes something mentally internal, say, from what makes it epistemically internal. Something can be mentally internal, yet epistemically external, as I am about to argue.)

5. Since I am discussing only justification *conditions* (see n. 2 above), not *theories*, deciding that some condition of your having justification is epistemically internal does not entail that justification as a whole is epistemically internal. (Presumably, though, if *all* the conditions of justification are epistemically internal, then justification is too.)

6. Strictly, this "in way *W*" is redundant. Some instance of *A* contributing in way *W* is *itself* an instance of *A*. (So, it will transpire, epistemic internalism asks that you appreciate not only *that A* is contributing to your justification, but *how* it is doing so. For if *A* contributes, it does so in *some way W*: *A*-contributing-in-way-*W* is therefore contributing too. Epistemic internalists will not *want* to admit that their view asks this of you, I presume (e.g. Audi 1988, 112.) But, as I argued in Section I above, they *should* do so. Either that, or they should become epistemic *ex*ternalists.)

7. Or, at least, if the latter aspect if epistemically external, then the former one will not be wholly internal. But I will assume that a condition (as against a theory: see n. 5 above) is either wholly internal or wholly external.

8. My assumption is based on the following plausibility consideration. Each member of the sequence *depends*, for its status as epistemically internal, on the next member. Moreover, this dependence is not merely identity. In general, your appreciating that *A* contributes to your being justified is not the same as your appreciating *that* the former appreciating contributes to your being justified.

9. And, in the spirit of n. 6 above, notice that if we let the appreciating-of-*A**-as-epistemically-internal (for some given instance *A** of *A*) be an instance of *A* itself, then this result recurs. In the very act of being appreciated as epistemically internal, the purportedly epistemically internal appreciating-of-*A**-as-epistemically-internal dies, becoming the appreciating-of-this-appreciating-of-*A**-as-epistemically-internal-as-itself-epistemically-internal. In other words, we encounter The Dilemma's First Horn.

10. All four of my examples will be more obviously applicable to non-inferential, than to inferential, justification. But the Dilemma still applies to both types of justification. It is simply easier to find historically important candidates for the epistemically internal among accounts of non-inferential justification. (One of the implications of the Dilemma is that this should *not* be easier, of course. Whether *A* is an immediate sensing or whether it is a logical relation between propositional contents, the Dilemma is equally relevant.)

11. See, for example, Fogelin 1987; Kripke 1982; McGinn 1984; Baker and Hacker 1984. (I discuss one aspect of the debate in my 1989.)

12. *S*'s being epistemically internal to you does require that you have a grasp of *S* (of *S*'s contributing to your justification). But is this grasp private, in that no one else *has* it? If not, we might argue that no one else could know of *S*'s justificatory role in the *way* you can, and hence that *S* (or its justificatory role) is in that sense private to you. In this paper I remain agnostic as to whether your private grasp would be linguistic (a use of language, in order to be aware of *S*'s justificatory role, which no one else could employ as you do). Hence I am not *assuming* that Wittgenstein's private language argument implies the Dilemma. Again, my conclusion is that *if S*'s being epistemically internal to you implies your using a private language to grasp *S*'s justificatory role, *then* Wittgenstein's private language argument applying to you is one aspect of the Dilemma applying to you.

13. For Russell, a logical particular is complete and self-subsistent (in Frege's terms, it is saturated). But for it to be epistemically internal you have to treat it as *in*complete, as *un*saturated, as needing to be *appreciated* by you, as needing

to be described by you in a special way. And to treat something complete as incomplete is to deny its being what it is.

14. It is not clear that Sellars is offering just one criticism of that view. I will outline one part of his thinking, a part which is particularly relevant to my discussion.

15. Ayer's own worry, in the second indented passage I quote from him, questions whether the basic proposition could even be *non*-epistemically internal (by being thought, for example, and hence being mentally, though not epistemically, internal). He seemingly believes that the basic proposition would be eliminated from the *entire* state of affairs, not just the *epistemic* state of affairs, since he is apparently implying that there can *be* no basic proposition for you to believe. I will not address that issue here, though.

BIBLIOGRAPHY

Alston, W. P. 1986. "Internalism and Externalism in Epistemology." *Philosophical Topics* 14: 179-221.

Audi, R. 1988. *Belief, Justification, and Knowledge*. Wadsworth.

Ayer, A. J. 1946. *Language, Truth and Logic* (second edn). Pelican.

Baker, G. P. and Hacker, P. M. S. 1984. *Scepticism, Rules and Language*. Blackwell.

BonJour, L. 1985. *The Structure of Empirical Knowledge*. Harvard University Press.

_____1987. "Nozick, Externalism, and Skepticism." In S. Luper-Foy (ed.), *The Possibility of Knowing*, at 297-313. Rowman & Littlefield.

Chisholm, R. M. 1989. *Theory of Knowledge* (third edn). Prentice-Hall.

Fogelin, R. J. 1987. *Wittgenstein* (second edn). Routledge & Kegan Paul.

Fumerton, R. 1988. "The Internalism/Externalism Controversy." *Philosophical Perspectives* 2: 443-459.

Goldman, A. I. 1979. "What is Justified Belief?" In G. S. Pappas (ed.), *Justification and Knowledge*, at 1-23. D. Reidel.

Hetherington, S. C. 1989. "Kripke and McGinn on Wittgensteinian Rule-Following." Currently submitted.

Kripke, S. A. 1982. *Wittgenstein on Rules and Private Language*. Harvard University Press.

McGinn, C. 1984. *Wittgenstein on Meaning*. Blackwell.

Russell, B. 1918. "The Philosophy of Logical Atomism." Reprinted in D. Pears (ed.), *Russell's Logical Atomism*, 1972, at 31-142. Fontana.

Pears, D. 1981. "The Function of Acquaintance in Russell's Philosophy." *Synthese* 46: 149-166.

Sellars, W. F. 1963. "Empiricism and the Philosophy of Mind." In *Science, Perception, and Reality*, at 127-196. Routledge & Kegan Paul.

Wittgenstein, L. 1958. *Philosophical Investigations*. (Trans. G. E. M. Anscombe.) Blackwell.

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 27, Number 3, July 1990

ON "WHY IS THERE SOMETHING RATHER THAN NOTHING?"

Martin Kusch

THIS paper does not try to resolve the question mentioned in the title; instead, it will argue that Robert Nozick's and Nicholas Rescher's answers to the question are unsatisfactory as they stand, and that both Martin Heidegger's and David Lewis's dismissals of the issue can be countered by raising doubts about their respective theoretical frameworks which lead to these dismissals.

2. Since Rescher's treatment of our title question in his recent book *The Riddle of Existence* (1984) surpasses the rival accounts in clarity and lucidity, it deserves to be treated first. Building on his *A Theory of Possibility* (1975) Rescher reformulates the title question as "Why should it be that the actually existing world is one of the non-empty ones...?" (1984: 6), a question that I shall shorten to "Why is the actual world non-empty?." Rescher's answer can be presented in the form of the following argument:

P1: There must be one—and only one—actual world.

P2: In order to be actualizable, a possible world must be really possible.

P3: Really possible worlds are non-empty.

C: The actual world must be non-empty.

Rescher supports P1 with the idea that given an inventory of all jointly exclusive and exhaustive distinct possibilities "one or another of them *must* obtain in the 'logical' nature of things" (Rescher 1984: 24). P2 is backed by the Leibnizian distinction within the realm of possibility, i.e., the distinction between logical, metaphysical (real), and physical possibility. To be really possible, for Rescher, is to be compatible with so-called "protophysical laws of nature," laws which "do not represent the behavioral dispositions of existents, but rather the *preconditions* to which something must conform if it is to become an existent at all. Such laws are not immanent in things but transcend their particular nature. They are 'laws of nature' alright, but in the rather special way of being laws *for* nature—laws that set preconditions upon the realizability of possibilities" (*ibid.*, p. 27).

Rescher holds that these laws explain the existence of a non-empty world since they rule out the empty world from among the really possible ones. Rescher regards the field equations of Einstein's General Theory of Relativity (GTR) as the "most plausible candidates" for protophysical laws. He suspects that it could perhaps be shown "that the only ultimately viable solutions to those equations are existential solutions...The cosmic equations would be such as to constrain existence in nature: they admit of no empty states and only allow for nonvacuous solutions" (*ibid.*, p. 34).

3. In order to gain a critical perspective on Rescher's suggestion, let us first turn to his rendering of our title question as "Why is the actual world non-empty?." Although this reformulation is backed by P1, at least Rescher's forerunner in searching for an answer to our question, and the philosopher in whose spirit Rescher claims to proceed, that is, Leibniz, seems to have posed the question rather as "Why is any world actual?." In his opuscule "On the Radical Origin of Things," Leibniz argues that it needs God to make any world actual (Leibniz 1697: 349). What makes the Leibnizian formulation attractive is that it does not force us to speak of nothingness as "the existence or actuality of an empty world." This latter expression sounds awkward because it forces us to make an existential claim with respect to a world where in fact we do not want to make any.

Of course, Leibniz's question causes trouble, too. Not only are we, in accepting it, forced to find a way around Rescher's arguments in support of P1 but we are also obligated to explain how we can account for the possibility of there being no actual world. When

no world is actual, then the possibility of there being no world is actual. How can we account for this actuality? To be sure, Leibniz could give an answer to this question by placing unactualized possible worlds, or rather their representations, in God's intellect (1697: 349). But this way of dealing with the question not only leads to the further problem of how we can reconcile God's existence with nothingness, it also makes far too strong metaphysical and/or theological assumptions.

The reason why we might nevertheless want to employ Leibniz's rather than Rescher's question is not only the strangeness of the existential claim "There exists an empty world," but also the further observation that Rescher's formulation potentially leads us into begging the question by thinking of the empty world as being somehow *just one more* possible world, as being *just like any other* possible world. To underline the fact that this danger is no mere fiction, we need to turn to P2 and P3 of Rescher's argument.

Recall that Rescher centrally relies on his idea that the set of protolaws constrains the set of potentially actual worlds, and that the same set of protolaws excludes the empty world from this set. But this seems to be an argument by stipulation as long as Rescher does not present us with convincing support for at least the following three claims, all of which are presupposed rather than argued for in his treatise:

> (CL 1) There is only one set of protolaws that defines the set of really possible worlds.

Why are we not allowed to assume that one subset of possible worlds is constrained by protolaws *a*, *b*, *c*, and that another distinct subset of possible worlds is constrained by protolaws *d*, *e*, *f*? Must we assume that the empty world is ruled out from actualization by any set of protolaws?

> (CL 2) In order to be actual the empty world must be really possible.

Why are we not allowed to say that in order to be really possible, a world must *either* fall within the domain of really possible worlds—this being a constraint on non-empty worlds—*or* else be empty?

> (CL 3) In order to be actual the empty world would have to be really possible *in the same way* as non-empty worlds have to be really possible in order to be candidates for actuality.

Rescher presupposes CL 3 because he tacitly equates possible worlds with possible natures. What makes this equation beg the question is his further notion that all possible *natures* are constrained by *GTR*. But what seems to be the crucial issue is not whether all *natures* are constrained, but whether an empty world, empty in a logical sense, can be called a nature at all. If the latter is denied, then the empty world might well be said to fall within the set of really possible worlds as defined by Rescher. We could then say that the empty world, precisely because it is no empty *nature*, is only vacuously constrained by protolaws; where there are no things, no states of affairs, no space and no time, that is, where there is no *nature*, constraints, whether physical or metaphysical, can hardly be violated.

4. For philosophers who reject the Rescherian notion of an empty world and who in addition reject the Leibnizian idea that no world needs to be actual, both Rescher's and Leibniz's reformulation of our title question must no doubt appear unintelligible. In fact, for them it is a logical necessity that something or other exists (at least when remaining committed to the possible worlds idiom). Proponents of this view will thus regard nothingness as a logical impossibility. Obviously, if they combine this double rejection with the notion of actuality as an indexical (an account first systematically developed and incorporated into modal metaphysics by David Lewis), they will have an even stronger reason to dismiss the question. Within Lewis's framework, one can ask neither the Leibnizian nor the Rescherian question since on his view every world is non-empty and actual "for itself" (as Hegel would have put it). Lewis thus has to take our title question as meaning "Why is this world actual and non-empty?," to which the answer can only be "Because all worlds are non-empty and actual for themselves, and because this world is actual, i.e., this worldly, for you."

Lewis's own formulation of the rebuttal is this: "'Why is there something rather than nothing?'—'If there were nothing, you wouldn't be here to ask the question.' Ask a silly question, get a silly answer..." (Lewis 1983: 23). Here Lewis does not make direct use of the idea that there is no empty world. Rather he seems to take the title question as "Why isn't the empty world actual (i.e., this worldly)?," his answer being basically that "Because then this world would not be this-worldly for you. The fact that *you* refer to

this world as *this-worldly* already presupposes that you *are* within this world, and that this world is thus not empty."

5. But certainly, this answer sounds odd and "too good to be true." In his recent book *On the Plurality of Worlds* (1986), Lewis admits once more—as he has already done in *Counterfactuals* (1973: 86)—that his modal realism "*does* disagree, to an extreme extent, with firm common sense opinion about what there is" (1986: 133): "When modal realism tells you—as it does—that there are uncountable infinities of donkeys and protons and puddles and stars, and of planets very much like Earth, and of cities very like Melbourne, and of people very like yourself,...small wonder if you are reluctant to believe it" (Lewis 1986: 133). Yet Lewis claims that believing all of this, despite our initial reluctance, is just the price we have to pay in order to enter the "philosophers' paradise" (1986: 1) of modal realism.

From the standpoint of our title question, however, modal realism does not look very much like a paradise. In fact we might even argue, via *modus tollendo tollens*, that if all we can say—based on Lewis's premisses—about our issue is "If there were nothing, you wouldn't be here to ask the question," then clearly there must be a flaw somewhere in these premisses. Put differently: What a strange kind of paradise for philosophers where their problems receive these kinds of answers! Obviously in cases where a theoretical framework reduces one of our questions to silliness, we have *prima facie* at least as much reason to be skeptical about the given theoretical framework as we have reason to accept the dismissal of our problem. Our decision as to which alternative we choose is of course dependent upon the weight we give to the respective question and to the success the theoretical framework has had in dealing with previous problems. Therefore all we can conclude with respect to Lewis's modal realism is this: Those of us who feel that the title question is a genuine one and who do not have independent overwhelming grounds for being or remaining modal realists might draw (additional) support for being opposed to modal realism from its inability to save our title question.

6. Let us once more return to Lewis's own rebuttal of our title query. In his argument, Lewis does not make use of his thesis that there is no empty world. Might we thus not modify his theory so that it allows for the possibility of the empty world? This is in fact what Robert Nozick does in *Philosophical Explanations* (1981). What Nozick finds intriguing about this modified conception of modal realism is that it allows for an "egalitarian theory" with respect to our title question. Nozick feels that the question "Why is there something rather than nothing?" is "inegalitarian" in that it implies that it is the existence of something (and not nothing) that stands in need of explanation. An egalitarian theory, instead, "will treat all possibilities on a par [and] one way to do this is to say that all possibilities are realized" (1981: 128). Nozick's solution: the possibility of nothing is realized in one world, and the possibility of something is realized in all other worlds: "Why is there something rather than nothing? There isn't. There's both" (1981: 130).

7. I have two reservations about this "solution." The first is but a variation of the often pronounced uneasiness with Lewis's modal realism and the counterpart theory that comes with it. Assume Barney will have his test results by tomorrow. Fred asks: "Why will Barney get a B rather than an A tomorrow?" Robert answers: "You shouldn't put it that way Fred. That is an inegalitarian question. Why will Barney get an A rather than a B tomorrow? He won't. He'll get both."—Moral: Being egalitarian *and* a modal realist makes you give funny answers to serious questions.

My second reservation is that I share Lewis's view that given his notion of possible worlds, we cannot make sense of an empty world. For Lewis a world is "a maximal mereological sum of spatiotemporally interrelated things" (1986: 73) and something that "by and large" is "concrete" (1986: 86). Nozick fails to give us an account of how he is prepared to modify Lewis's notion of world so that it allows for an empty world. I doubt that this could be done without giving up other crucial ingredients of modal realism, such as the indexical interpretation of actuality.

8. However, the argument just presented is only one of three different lines of thought that Nozick presents in his inventive essay "Why is there something rather than nothing?" (1981: 115-164). Moreover, all three can be naturally related to Heidegger's thought.

Nozick's Lewis-style approach to his and our query can be related to Heidegger only *via negationis*. Heidegger rejects talk of worlds in the plural, claiming that this talk turns worlds into ob-

jects. The idea is that "'world' is something within which one can live (and one cannot live within an object)" (Kusch 1989:296). We might spell out this idea—as was once suggested to me by Calvin Normore—by saying that to conceive of worlds as concrete objects is to use the notion of "concrete object" in a highly unusual and obscure way. It certainly is part of our usual concept of concrete objects that we can *point* to them. But in the case of a world, e.g. our world, we can only "wave our arms about in a vague way" (Prior & Fine 1977: 92).

9. Another, in fact Nozick's first way of dealing with our problem, is weakly related to Heidegger since—as Nozick observes himself—the way the idea is formulated at least *sounds* Heideggerian. The suggestion is that there is something rather than nothing because just like in the Beatles' cartoon *The Yellow Submarine* where a vacuum cleaner first sucks up all other things and finally itself, "thereby producing with a pop a brightly colored variegated scene," so also an initial "nothingness force" might have sucked up itself, "nothinged itself, thereby producing something" (1981: 123). Since I find this suggestion not only quite foreign to Heidegger but also rather obscure—if there already was the vacuum cleaner, or nothingness force, there already was something—let me move on straight away to Nozick's third way.

10. Nozick's third way of treating something and nothing (1981: 150-164) does not seem to be even intended to answer our title question. He queries whether the dichotomy "exist *versus* does not exist" is exhaustive, whether this dichotomy is grounded in a presupposition that makes it possible, and finally, whether we can conceive of an area falling outside of this opposition. Nozick ultimately refers to mystical experience in order to answer all three questions positively.

Instead of following him in doing so, and instead of criticizing him for doing so—this has already been done by Rescher (1984: 5)—let me rather turn to Heidegger and use Nozick's query as a starting point for a brief summary of Heidegger's position *vis-à-vis* "Why is there being (*Seiendes*) rather than nothing?" (1953: 1).

Heidegger indeed holds with Nozick that this question is based on a presupposition and thus the alternative "being *versus* nothing" is not exhaustive. To see what Heidegger considers the presupposition to be, note that he calls this question "the basic question of metaphysics" (1953: 13) and that instead of answering it he proceeds to answer the "preliminary question" (*Vorfrage*), "*Wie steht es um das Sein*?" (1953: 25). Now metaphysics for Heidegger tacitly worked and works on the presupposition that Being means "presence" (*Anwesenheit*). That is to say, metaphysics is based on the model of the explicitly identified object, on the model of the object that is explicitly picked out and "made present" in perception, on the object that that is at the focus of perceptual attention. For metaphysics—roughly speaking—beings come one by one, or piece by piece, and not only can each one be isolated and turned into an object for the investigating subject, but even their total sum can be treated in the same manner. Even with this rough sketch, we can see why our title question is metaphysical: only a thinking based on the presupposition of Being as presence can talk about being (*Seiendes*), or beings, as a whole, and set this whole apart as one possibility from the possibility of nothing.

Now the early Heidegger thought that that he could pinpoint what this presupposition rules out: the Being (*Sein*) of the human being (*Dasein*), and the Being of "equipment" (*Zeug*). The mode of Being of both is such that to subsume them indifferently under the category of "being" (*Seiendes*) or "something," i.e. to subsume them indifferently under "present(-to-hand)" something—a subsumption that the metaphysical basic question involves—is to commit something like a category mistake.

The later Heidegger seems to be more pessimistic as to the question whether we can do anything more than just hint—by poetic language—at what lies beyond the metaphysical presupposition. This pessimism is based on the view—already tacitly at work in the period of *Being and Time*—that Being as the transcendental condition of our language cannot be spoken about in anything but a tautological fashion. Being, language, and world are just one universal medium of meaning, and none of these can be spoken *about*: all we are able to do, is to speak *from within* them. To use an expression from *Being and Time* we might say that Being, language and world form one "closed whole" (1962: 103). We can only speak about things within the endless horizon of our world, a horizon opened up for us by our understanding of Being, while we can speak neither about language—Heidegger regards metalanguage as an

abuse of language—, nor about Being, nor about the world or being (*Seiendes*) as a whole. Even though Heidegger does not put it in so many words, we might conclude that the basic metaphysical question is to be dismissed: the whole world cannot even be hypothetically assumed to be non-existing or empty.

11. Heidegger's dismissal of our title question is certainly more difficult to counter than Lewis's, since Heidegger's rejection is based on an all-embracing theory of the whole western philosophical tradition. Needless to say, a criticism of this theory lies beyond the scope of this paper. I shall therefore confine myself to two suggestions.

First, Heidegger has not shown convincingly—perhaps he has not even wanted to claim—that subsuming *Dasein* and *Zeug* under the category of beings-in-general or something, always and inevitably involves a mistake or an inauthentic procedure. To be sure, when trying to clarify *Dasein's* and *Zeug's* mode of Being, we must not employ this subsumption. But may not *Dasein*, while being well aware of its primordial mode of Being, still, for the purpose of some inquiry, look upon itself (himself, herself, ourselves) as just one present-to-hand being among others? Yet if we allow for this possibility, then we must also allow *Dasein* to ask the basic metaphysical question.

Second, as I have shown elsewhere, Heidegger's conception of the closed whole, i.e. the one-world-one-language conception, has a number of uncomfortable consequences such as the ineffability of language-world relations, the denial of metalanguage and possible worlds, semantical Kantianism, (linguistic) relativism, and even a rather strong determinism (Kusch 1989: 148-228). At least those of us who believe that we have independent good reasons for rejecting these consequences also have some grounds for being more than just skeptical about the premisses from which such unhappy corollaries flow. Yet once these premisses stand in doubt, the shadow which they throw over our title question shrinks quickly.

12. To conclude, while Rescher and Nozick fail to provide a compelling answer to our title question, so do Lewis and Heidegger in trying to dismiss it. We are left with nothing but negative results. If I do not find this fact overly disturbing, it is perhaps because I believe that at least in philosophy the nothingness vacuum cleaner has some purpose: while it doesn't produce worlds, it still keeps a conceptually productive debate going. And that's more than nothing!

University of Oulu, Finland

Received October 10, 1989

BIBLIOGRAPHY

Heidegger, Martin (1953): *Einführung in die Metaphysik* (Tübingen, Niemeyer).

Heidegger, Martin (1962): *Being and Time*, trans. by John Macquarrie & Edward Robinson (Oxford, Basil Blackwell).

Kusch, Martin (1989): *Language as Calculus vs. Language as the Universal Medium. A Study in Husserl, Heidegger and Gadamer* (Dordrecht, Kluwer).

Kusch, Martin (1990): "Heidegger on 'Why Is There Something Rather Than Nothing?'," forthcoming in *Acta Philosophica Fennica.*

Leibniz, Gottfried Wilhelm von (1697): "On the Ultimate Origin of Things," in: *Leibniz Selections*, ed. by Philip P. Wiener (New York, Charles Scribner's Sons).

Lewis, David (1983): *Philosophical Papers* (Oxford, Basil Blackwell).

Lewis, David (1986): *On the Plurality of Worlds* (Oxford, Basil Blackwell).

Nozick, Robert (1981): *Philosophical Explanations* (Oxford, Clarendon Press).

Prior, Arthur and Kit Fine (1977): Worlds, Time and Selves (Amherst, University of Massachusetts Press).

Rescher, Nicholas (1975): *A Theory of Possibility* (Oxford, Basil Blackwell).

Rescher, Nicholas (1984): *The Riddle of Existence* (Lanham, University Press of America).

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 27, Number 3, July 1990

The Editor's Page

WHERE WISE MEN FEAR TO TREAD

It is a cheerless—though doubtless unsurprising—fact that trained philosophers are just as captivated by the political fashions and frenzies of their place and time as other people are. But since they presumably realize the pitfalls involved, it is somewhat more distressing that philosophers are not epistemically more cautious about voicing their opinions and more restrained by considerations of common sense in launching them into print. A few illustrations show that these strictures are not unjustified.

1. Ralph Barton Perry on the wickedness of the Germans in the era of World War I: "Being convinced, for example, that the state has a divine mission and is entitled to a dominion proportioned to its power, the German is not deterred by the protests of those who stand in the way. Or having once adopted a certain theory of warfare, and reconciled it with this higher law of the state, the German is not rendered the least irresolute by the incidental sufferings which he inflicts. In carrying out his preconceived ideas, the German is also peculiarly able to harden himself against moral tradition and the opinion of mankind." (*The Present Conflict of Ideas* [New York: Longmans, Green; 1918], p. 401.)

2. John Dewey on Soviet Communism after a brief visit to the U.S.S.R. in the late 1920's: "[T]he most significant aspect of the change in Russia is psychological and moral, rather than political.... I had the notion that socialistic communism was essentially a purely economic scheme.... That the movement in Russia is intrinsically religious was something I had often heard and that I supposed I understood and believed. But when fact to fact with actual conditions, I was forced to see that I had not understood it at all...for this failure there were two causes.... One was that, never having previously witnessed a widespread and moving religious reality, I had no way of knowing what it actually would be like. The other was that I associated the idea of Soviet Communism, as a religion, too much with intellectual theology, the body of Marxian dogmas, with its professed economic materialism, and too little with a moving human aspiration and devotion. As it is, I feel as if for the first time I might have some inkling of what may have been the moving spirit and force of primitive Christianity." *Impressions of Soviet Russia*, reprinted in New York in 1966, pp. 104-105.

3. Simone de Beauvoir on the use of informers in Mao's China: "Urging people to vigilance, the government does indeed exhort them to report the counterrevolutionary activities whereof they may have cognizance; but we must not forget that these activities consist in arson, the sabotage of bridges and dikes, in assassinations...." (*The Long March* [Cleveland and New York, 1958], pp. 388-389.)

4. Jean Paul Sarte's reaction to the Gulag archipelago in the wake of a visit to Stalin's U.S.S.R.: "As we were neither members of the party nor avowed sympathizers, it was not our duty to write about Soviet labor camps; we were free to remain aloof from the nature of this system, provided no events of sociological significance occurred." (As quoted in Walter Laqueur and G.L. Mosse, ed's, *Literature and Politics in the Twentieth Century* [New York, 1967], p. 25.)

5. Noam Chomsky discounting "the tales of Communist atrocities" in the killing fields of the Pol Pot regime maintains that: "Executions have numbered at most in the thousands; and that these were localized in areas of limited Khmer Rouge influence and unusual peasant discontent, where brutal revenge killings were aggravated by the threat of starvation resulting from the American destruction and killing." (Noam Chomsky and E.S. Herman "Distortions at Fourth Hand," *The Nation*, June 25, 1977; p. 791.)

There is much to be said for Tayllerand's injunction: *pas trop de zèle*. In writing political commentary as in writing love letters, opinions expressed amidst the warm glow of enthusiasm often look problematic after the embers cool. That the near unavoidable recognition of this fact does not influence *philosophers* more does not speak well for their capacity to absorb the lessons of their own discipline. In this light, it is debatable whether it is the world or the discipline of philosophy that has been the worst served by Marx's insistence that philosophers should seek to change the world rather than to understand it.

BOOKS RECEIVED

Arendt, Hannah, *Lectures on Kant's Political Philosophy* (Chicago: The University of Chicago Press, 1989), 174 pp., $9.95.

Audretsch, Jürgen, and Klaus Mainzer, *Vom Anfang der Welt: Wissenschaft, Philosophie, Religion, Mythos* (München: C. H. Beck, 1989), 228 pp.

Baird, Robert M., and Stuart E. Rosenbaum, *Euthanasia: The Moral Issues* (Buffalo, NY: Prometheus Books, 1989), 182 pp., $11.95.

Bernhardt, Jean, *Hobbes* (Paris: Presses Universitaires de France, 1989), 126 pp.

Brabeck, Mary M., *Who Cares? Theory, Research, and Educational Implications of the Ethic of Care* (New York: Praeger Publishers, 1989), 250 pp., $45.00.

Chambre, Cureau de la, *Traite de la Connaissance des Animaux*; reprint (Librairie Artheme Fayard, 1989), 368 pp.

Colander, David, and A. W. Coats, *The Spread of Economic Ideas* (Cambridge: Cambridge University Press, 1989), 262 pp.

Descartes, Rene, *The Passions of the Soul*, translated and annotated by Stephen Voss (Indianapolis: Hackett Publishing Company, 1989), 183 pp., $22.50 cloth, $4.95 paper.

Desmond, William, *Hegel and His Critics: Philosophy in the Aftermath of Hegel* (NY: State University of New York Press, 1989), 242 pp., $44.50 cloth, $14.95 paper.

Donaldson, Thomas, *The Ethics of International Business* (NY: Oxford University Press, 1989), 196 pp., $24.95.

Downing, Christine, *Myths and Mysteries of Same-Sex Love* (NY: The Continuum Publishing Co., 1989), 317 pp., $22.95.

Earman, John, *World Enough and Space-Time: Absolute versus Relational Theories of Space and Time* (Cambridge, MA: The MIT Press, 1990), 233 pp., $25.00.

Farrer, Austin, *Faith and Speculation* (Edinburgh: T & T Clark, 1988), 192 pp., $16.95.

Faye, Jan, *The Reality of the Future: An Essay on Time, Causation and Backward Causation* (Odense, Denmark: Odense University Press, 1989), 321 pp.

Fraser, Nancy, *Unruly Practices: Power, Discourse, and Gender in Contemporary Social Theory* (Minneapolis, MN: University of Minnesota Press, 1989), 201 pp., $14.95.

Griffin, David Ray, and Huston Smith, *Primoridal Truth and Postmodern Theology* (Albany, NY: State University of New York Press, 1989), 216 pp., $39.50 cloth, $12.95 paper.

Habermas, Jürgen, *The New Conservatism: Cultural Criticism and the Historians' Debate*, edited and translated by Shierry Weber Nicholsen (Cambridge, MA: The MIT Press, 1990), 270 pp., $19.95.

Heil, John, *Cause, Mind, and Reality: Essays Honoring C. B. Martin* (Dordrecht: Kluwer Academic, 1989), 300 pp.

Klein, Jacob, *A Commentary on Plato's Men* (Chicago: The University of Chicago Press, 1989), 256 pp., $14.95.

Marcel, A. J. and E. Bisiach, *Consciousness in Contemporary Science* (Oxford: Clarendon Press, 1988), 405 pp., $75.00.

MacGregor, Geddes, *Dictionary of Religion and Philosophy* (NY: Paragon House, 1989), 712 pp., $35.00.

Machan, Tibor R., *Liberty and Culture: Essays on the Idea of a Free Society* (Buffalo, NY: Prometheus Books, 1989), 288 pp., $19.95.

McCarthy, Michael H., *The Crisis of Philosophy* (New York: State University of New York Press, 1989), 383 pp., $59.50 cloth, $19.95 paper.

Pollock, John, *How To Build a Person* (Cambridge, MA: The MIT Press, 1990), 189 pp., $22.50.

Posner, Michael I., *Foundations of Cognitive Science* (Cambridge, MA: The MIT Press, 1990), 880 pp., $45.00.

Pruzan, Elliot R., *The Concept of Justice in Marx* (New York: Peter Lang Pub., Inc., 1989), 238 pp., $37.00.

Schueler, G. F., *The Idea of a Reason for Acting* (Lewiston, NY: The Edwin Mellen Press, 1989), 113 pp., $39.95.

Seidman, Steven, ed., *Jürgen Habermas on Society and Politics: A Reader* (Boston, MA: Beacon Press, 1989), 324 pp., $16.95.

Strawson, Galen, *Freedom and Belief* (New York: Oxford University Press, 1987), 339 pp., $39.95.

Winter, Gibson, *Community and Spiritual Transformation: Religion and Politics in a Communal Age* (NY: Crossroad Publishing Co., 1989), 135 pp.

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 27, Number 4, October 1990

ON NATURALIZING EPISTEMOLOGY

Robert Almeder

I. INTRODUCTION

THERE are three distinct forms of naturalized epistemology. The first form asserts that the only legitimate questions about the nature of human knowledge are those we can answer in natural science. So described, naturalized epistemology is a branch of natural science wherein the questions asked about the nature of human knowledge make sense only because they admit of resolution under the methods of such natural sciences as biology and psychology. Characterized in this way, naturalised epistemology consists in empirically describing and scientifically explaining how our various beliefs originate, endure, deteriorate or grow. Unlike traditional epistemology, this form of naturalized epistemology does not seek to determine whether the claims of natural science are more or less justified. For this reason, it is not "normative" in the way traditional epistemology is normative. Not surprisingly, this first form of naturalized epistemology regards traditional "philosophical" questions about human knowledge, questions whose formulation and solution do not emerge solely from the practice of natural science, as pointless. Accordingly, this first form of naturalized epistemology seeks to *replace* traditional epistemology with the thesis that while we certainly have scientific knowledge, and whatever norms are appropriate for the successful conduct of natural science, we have no philosophical theory of knowledge sitting in judgment over the claims of natural science to determine whether they live up to a philosophically congenial analysis of justification or knowledge. As we shall see shortly, the classical defense of this first form of naturalized epistemology appears in Quine's "*Naturalized Epistemology*".[1]

The second form of naturalized epistemology seeks less to *replace* traditional epistemology than it does to *transform* and supplement it by connecting it with the methods and insights of psychology, biology and cognitive science. In *Epistemology and Cognition*, for example, Alvin Goldman has argued for this second form which allows for traditionally normative elements but is "naturalized" for the reason that the practitioners of natural science, especially biology and psychology, will have the last word on whether anybody knows what they claim to know. For Goldman, although defining human knowledge and other epistemic concepts is legitimately philosophical and traditionally normative, whether anybody knows what they claim to know, and just what cognitive processes are involved, is ultimately a matter we must consign to psychologists or cognitive scientists. Unlike the first form of naturalized epistemology, this second form allows traditional epistemology to sit in judgment on the claims of natural science but the judgment must be made by the practitioners of natural science using the methods of natural science.

The third distinct form of naturalized epistemology simply insists that the method of the natural sciences is the only method for acquiring a proper understanding of the nature of the physical universe. On this view, natural science, and all that it implies, is the most epistemically prvileged activity for understanding the nature of the physical world. Adopting this last form of naturalized epistemology is, however, quite consistent with rejecting both of the above forms of naturalized epistemology. This third form is quite compatible with traditional epistemology because it does not seek to *replace* traditional epistemology in the way that the Quine thesis does; nor does it seek to *transform* traditional epistemology by turning the question of who knows what over to psychologists and cognitive scientists in the way that the Goldman thesis does.

At any rate, the most currently pervasive and chal-

lenging form of naturalised epistemology is the radically anti-traditional, anti-philosophical thesis offered originally by Quine and recently defended by others. So, in the next few pages we shall focus *solely* on the Quinean thesis and five distinct arguments recently offered in defense of it. Along the way, we will discuss various objections to such a naturalized epistemology, objections proponents of the thesis have recently confronted. Unfortunately, because space is here limited, we will not examine the second and distinct form of naturalised epistemology offered by Alvin Goldman. To do so would involve a long discussion of the merits of the reliabilist theory of justification upon which Goldman's type of naturalism squarely rests.[2]

Finally, the modest conclusion of this paper is that there is no sound argument available for the Quinean form of naturalized epistemology. The immodest conclusion is that any argument proposed for the thesis will be incoherent, and that consequently there is no rational justification for anybody taking such a naturalistic turn.

II. Quine's Argument

In "Epistemology Naturalized", Quine begins his defense of naturalized epistemology by asserting that traditional epistemology is concerned with the foundations of science, broadly conceived. As such, it is supposed to show how the foundations of knowledge, whether it be the foundations of mathematics or natural science, reduce to certainty. In short, showing how certainty obtains is the core of traditional epistemology, and this implies that the primary purpose of traditional epistemology is to refute the Cartesian sceptic whose philosophical doubts over whether we can attain certainty has set the program for traditional epistemology.

But, for Quine, traditional epistemology has failed to refute the sceptic, and will never succeed in refuting the sceptic. Mathematics reduces only to set theory and not to logic; and even though this reduction enhances clarity it does nothing by way of establishing certainty because the axioms of set theory have less to recommend them by way of certainty than do most of the mathematical theorems we would derive from them. As he says:

> Reduction in the foundations of mathematics remains mathematically and philosophically fascinating, but it does not do what the epistemologist would like of it: it does not reveal the ground of mathematical knowledge, it does not show how mathematical certainty is possible. (p.71).

Moreover, mathematics aside, the attempt to reduce natural knowledge to a foundation in the certainty of statements of sense experience has also failed miserably. Common sense about sensory impressions provides no certainty. And, when it comes to justifying our knowledge about truths of nature, Hume taught us that general statements and singular statements about the future do not admit of justification by way of allowing us to ascribe certainty to our beliefs associated with such statements. For Quine, the problem of induction is still with us; "The Humean predicament is the human predicament." (p.72). As Quine sees it, Hume showed us quite clearly that any attempt to refute the sceptic by uncovering some foundation of certainty associated with sense statements, whether about sense impressions or physical objects, is doomed equally to failure.(p.72)

This last consideration is crucial because as soon as we accept Quine's rejection of the analytic/synthetic distinction in favor of only synthetic propositions, Hume's argument casts a long despairing shadow over our ever being able to answer the sceptic because such propositions could never be certain anyway. The conclusion Quine draws from all this is that traditional epistemology is dead. There is no "first philosophy". There are no strictly philosophical truths validating the methods of the natural sciences. Nor can we validate in any non-circular way the methods of the natural sciences by appeal to psychology or the methods of the natural sciences. As he says, "If the epistemologist's goal is validation of the grounds of empirical science, he defeats his purpose by using psychology or other empirical science in the validation." (pp.75-76). We may well have justified beliefs based upon induction, but we cannot have any justified belief that we can have justified beliefs based upon induction. Accordingly, if epistemology is to have any content whatever, it will seek to explain, *via* the methods of natural science, the origin and growth of beliefs we take to be human knowledge and natural science. Construed in this way, epistemology continues as a branch of natural science wherein the only meaningful questions are questions answerable in science by scientists using the methods of natural science. This

reconstrual of the nature of epistemology consigns the enterprise to a descriptive psychology whose main function is to describe the origin of our beliefs and the conditions under which we take them to be justified. On this view, all questions and all doubts are scientific and can only be answered or resolved in science by the methods of science. Philosophical discussions on the nature and limits of scientific knowledge, questions that do not lend themselves to resolution via the methods of natural science are simply a part of traditional philosophy that cannot succeed. What can we say about all this?

III. Response to Quine's Argument

In "The Significance of Naturalized Epistemology," Barry Stroud criticizes Quine's defense of naturalized epistemology.[3] After a brief description of Quine's position, Stroud argues that Quine is inconsistent for arguing *both* that there is no appeal to scientific knowledge that could non-circularly establish the legitimacy of scientific knowledge in the presence of the traditional epistemological sceptic, *and* in *Roots of Reference* that we should take seriously the project of validating our knowledge of the external world.[4] For Stroud, it was in *Roots of Reference* that Quine came to believe in the coherent use the resources of natural science to validate the deliverances of natural science. But that would be to countenance the basic question of traditional epistemology when in fact the thrust of Quine's thesis on naturalized epistemology is that such a question forms part of 'first philosophy' which is impossible. Apart from such an inconsistency, Stroud also argues that Quine's attempt to validate scientific inference fails. (p.81) Stroud's thesis here is that Quine attempts to offer a naturalized defense of science in " The Nature of Natural Knowledge,"[5] but the effort fails because, on Quine's reasoning, we can see how others acquire their beliefs but we are denied thereby any evidence of whether such beliefs are correct beliefs about the world. By implication, we have no reason for thinking our own beliefs are any better off. (p. 81) In commenting on Quine's defense, Stroud says:

> Therefore, if we follow Quine's instructions and try to see our own position as 'just like' the position we can find another "positing' or 'projecting' subject to be in, we will have to view ourselves as we view another subject when we can know nothing more than what is happening at his sensory surfaces and what he believes or is disposed to assert. (p. 81)

His point here is that when we examine how another's beliefs originate, we have no way to look beyond his positing to determine whether his beliefs are true or correct. In that position we never can understand how the subject's knowledge or even true belief is possible. Therefore, we never can understand how our own true beliefs are possible either. Stroud says:

> The possibility that our own view of the world is a *mere* projection is what had to be shown not to obtain in order to explain how our knowledge is possible. Unless that challenge has been met, or rejected, we will never understand how our knowledge is possible at all. (p.83)

> ...if Quine's naturalized epistemology is taken as an answer to the philosophical question of our knowledge of the external world, then I think that for the reasons I have given, no satisfactory explanation is either forthcoming or possible. (p. 83)

He goes on to conclude that if naturalized epistemology is *not* taken as an answer to the philosophical question of our knowledge of the external world, and if the question is a legitimate question (and Quine has not shown that it is not) then naturalized epistemology cannot answer the question:

> I conclude that even if Quine is right in saying that sceptical doubts are scientific doubts, the scientific source of these doubts has no anti-sceptical force in itself. Nor does it establish the relevance and legitimacy of a scientific epistemology as an answer to the traditional epistemological question. If Quine is confident that a naturalized epistemology can answer the traditional question about knowledge, he must have some other reason for that confidence. He believes that sceptical doubts are scientific doubts and he believes that in resolving those doubts we may make free use of all the scientific knowledge we possess. But if, as he allows, it is possible for the skeptic to argue by *reductio* that science is not known, then it cannot be that the second of those beliefs (that a naturalized epistemology is all we need) follows from the first.

> Until the traditional philosophical question has been exposed as in some way illegitimate or incoherent, there will always appear to be an intelligible question about human knowledge in general which, as I have argued, a naturalised epistemology cannot answer. And Quine himself seems committed at least to the coherence of

> that traditional question by his very conception of knowledge. (pp. 85-86)

Stroud's closing remark is that the traditional question has not been demonstrated as illegitimate, and Quine's attempt to resolve skeptical doubts as scientific doubts within science has failed. Moreover, for Stroud, apart from the question of whether Quine succeeded, his effort is predicated on the legitimacy of the traditional question of whether science provides us with knowledge of the external world.

Some naturalized epistemologists will probably disagree with Stroud's analysis and urge that Quine's attempt to validate scientific knowledge is misunderstood when construed as an attempt to establish first philosophy. Better by far that we read Quine as asserting that there is simply no way to validate the deliverances of science as more or less warranted. Whether this last response is adequate, we cannot now discuss.

At any rate, Quine has responded to Stroud with the following remarks:

> What then does our overall scientific theory really claim regarding the world? Only that it is somehow structured as to assure the sequences of stimulation that our theory gives us to expect....
>
> In what way then do I see the Humean predicament as persisting? Only in the fallibility of prediction: the fallibility of induction and the hypothetico-deductive method in anticipating experience.
>
> I have depicted a barren scene. The furniture of our world, the people and sticks and stones along with the electrons and molecules, have dwindled to manners of speaking. And other purported objects would serve as well, and may as well be said already to be doing so.
>
> So it would seem. Yet people, sticks, stones, electrons and molecules are real indeed, on my view, and it is these and no dim proxies that science is all about. Now, how is such robust realism to be reconciled with what we have just been through? The answer is naturalism: the recognition that it is within science itself, and not in some prior philosophy, that reality is properly to be identified and described.[6]

In reflecting on this response to Stroud, Ernest Sosa has been quick to note the incoherence involved in accepting science as the "reality-claims court coupled with the denial that it is anything but free and arbitrary creation."[7] Continuing his criticism of Quine, Sosa goes on to say:

> The incoherence is not removed, moreover, if one now adds:
>
> (Q1) What then does our overall scientific theory really claim regarding the world? Only that it is somehow so structured as to assure the sequence of stimulation that our theory gives us to expect.
>
> (Q2) Yet people, sticks, stones, electrons and molecules are real indeed.
>
> (Q3) [It]...is within science itself and not in some prior philosophy, that reality is properly to be identified and described.
>
> If it is within science that we settle, to the extent possible for us, the contours of reality; and if science really claims regarding the world only that it is so structured as to assure certain sequences of stimulation; then how can we possibly think reality to assume the contours of people, sticks, stones and so on?
>
> We cannot have it all three ways: (Q1), (Q2) and (Q3) form an incoherent triad. If we trust science as the measure of reality, and if we think there really are sticks and stones, then we can't have science accept only a world 'somehow so structured as to assure' certain sequences of stimulations or the like. Our science must also claim that there really are sticks and stones.
>
> What is more, if science really is the measure of reality it cannot undercut itself by saying that it really isn't, that it is only convenient 'manners of speaking' to guide us reliably from stimulation to stimulation. (p.69)

Sosa's criticism seems quite pointed. Morever, even if the critique offered by both Stroud and Sosa should turn out to be a misconstrual of Quine's position, there are other plausible objections we might raise to Quine's argument for naturalized epistemology. For one thing, it is obvious that Quine's argument itself for naturalized epistemology is a philosophical argument, which, *ex hypothesi,* should not count by way of providing evidence for the thesis of naturalized epistemology. Further, the thesis of naturalized epistemology is arguably fundamentally incoherent. It argues against there being a "first philosophy" by appealing to two premisses both of which are sound only if philosophical arguments about the limits of human knowledge are permissible and sound. The first premise consists in asserting that Hume's skepticism about factual knowledge is indeed established. Hume's thesis is certainly not empirically confirmable. The so-called "problem of induction" is a philosophical problem

based upon a certain philosophical view about what is necessary for scientific knowledge. It is certainly not a problem in natural science or naturalised epistemology. The second premise is the denial of the analytic/synthetic distinction; and that is a thesis largely resting on a philosophical argument about the nature of meaning. Such premisses only make sense within a commitment to the validity of some form of first philosophy and the legitimacy of traditional epistemology. Finally, there is the problematic premise that traditional epistemology has been exclusively concerned to establish the foundations of certainty in order to show that we have knowledge of the world. A close look at traditional epistemology, however, suggests that the primary concern is as much a matter of getting clear on, or (as Sosa has noted) understanding just what it *means* to know, and just what the concept of certainty relative to different senses of "knows" consists in, as it is a matter of validating knowledge claims or seeking the foundations of certainty (p. 50-51). Indeed, there is good reason to think that the primary concern of traditional epistemology is one of *defining* concepts of knowledge, certainty, justification and truth; and only then of determining whether anybody has the sort of certainty associated with the correct definition of knowledge. The history of epistemology is as apt to criticize the program of the Cartesian sceptic (and the definition of knowledge implied therein) as it is to accept it. Certainly, if the concept of knowledge and justification had been defined differently than Hume had defined them, Hume's predicament would never have occured. Traditional epistemology is probably as concerned with what it means for a belief to be certain, as it is with determining whether scientific knowledge is certain. With these few considerations in mind, and realizing that much more can be said on the issue, it would appear that Quine's defense of naturalized epistemology admits of a number of solid objections. Let us turn to a more recent and quite distinct defense of the Quinean thesis.

IV. The "Philosophy is Science" Argument

Among recent arguments for the Quine thesis, the argument offered by William Lycan in *Judgement and Justification* is quite different from Quine's.[8] Unlike the Quinean argument, it does not rest on the alleged failure of the analytic/synthetic distinction and upon the subsequent classification of all propositions as synthetic. Nor does it feed upon Quine's Humean argument that synthetic propositions cannot be justified from the viewpoint of a first philosophy and so, if epistemology is to continue, it can only be in terms of the deliverances of a descriptive psychology. What is the argument?

Lycan begins by characterising classical philosophy in terms of the deductivist model. He calls it "deductivism" and under this model philosophy gets characterised in a certain way:

> Philosophers arrive at conclusions that are guaranteed to be as indisputably true as the original premisses once the ingenious deductive arguments have been hit upon. This attitude has pervaded the rationalist tradition and survives among those who are commonly called 'analytic philosophers' in a correctly narrow sense of that expression. Of course the quality of self-evidence that the deductivist premises are supposed to have and to transmit to his conclusion has been variously described (as for example, analyticity, a prioricity, clarity and distinctness, mere obviousness, or just the property of having been agreed upon by all concerned.)....
>
> In any case, the deductivist holds that philosophy and science differ in that deductive argument from self-evident premises pervades the former but not the latter. (Some deductivists have held a particularly strong version of this view, identifying the philosophy/science distinction with the a priori/empirical distinction.) Now Quine (1960, 1963, 1970) has a rather special reason for rejecting the dichotomy between philosophy and science, or, to put the point more accurately, between philosophical method and scientific method. He rejects the analytic-synthetic distinction, and thus the proposal that there are two kinds of truths ("conceptual" or "a priori", and "empirical" or scientific.), one of which is the province of philosophy and the other the province of science....
>
> I side with Smart and Quine against the deductivist, but for what I think is a more fundamental and compelling reason, one that does not depend on the rejection of the analytic/synthetic distinction. If my argument is sound, then one can countenance that distinction and still be forced to the conclusion that the Smart/Quine methodological view is correct.
>
> Suppose we try to take a strict deductivist stance. Now, as is common knowledge, one cannot be committed (by an argument) to the conclusion of that argument unless one accepts the premisses. Upon being presented with a valid argument, I always have the option of denying

its conclusion, so long as I am prepared to accept the denial of at least one of the premisses.

Thus, every deductive argument can be set up as an inconsistent set. (Let us, for simplicity, consider only arguments whose premisses are internally consistent.) Given an argument *P*, *Q* ∴ *R* the cognitive cash value of which is that R follows deductively from the set of *P* and *Q*, we can exhaustively convey its content simply by asserting that the set (*P*, *Q*, not-*R*) is inconsistent, and all the original argument has told us, in fact, is that for purely logical reasons we must deny either *P*, *Q*, or not-*R*. The proponent of the original argument, of course, holds that *P* and *Q* are true; therefore, she says, we are committed to the denial of not-*R*, that is, to *R*. But how does she know that *P* and *Q* are true? Perhaps she has constructed deductive arguments with *P* and *Q* as conclusions. But, if we are to avoid regress, we must admit that she relies ultimately on putative knowledge gained nondeductively; so let us suppose that she has provided nondeductive arguments for *P* and *Q*. On what grounds then does she accept them? The only answer that can be given is that she finds each of *P* and *Q* more plausible than not-*R*, just as Moore found the statement "I had breakfast before I had lunch" more plausible than any of the metaphysical premises on which rested the fashionable arguments against the reality of time.

But these are just the sorts of considerations to which the theoretical scientist appeals. If what I have said here is (more or less) right, then we appear to have vindicated some version of the view that (1) philosophy, except for that relatively trivial part of it that consists in making sure that controversial arguments are formally valid, is just very high level science and that consequently (2) the proper philosophical method for acquiring interesting new knowledge cannot differ from proper scientific method. (pp. 116-118).

In defending this general argument, Lycan then responds to two basic objections which he offers against his own thesis. The first objection is that even if all philosophical arguments rest on plausibility arguments, the above argument has not established what is necessary, namely, that considerations that make for plausibility in science are the same considerations that makes for plausibility in philosophy. The second objection is that even if we were to establish as much, it would not thereby obviously follow that philosophy is just very high-level science. Philosophy and science might have the same method but differ by way of subject matter. (pp. 116-118). The core of Lycan's defense of his general argument consists in responding to the first objection. So, let us see whether the response overcomes the objection.

In response to the first objection, Lycan constructs the following argument:

P1. The interesting principles of rational acceptance are not the deductive ones (even in philosophy).

P2. There are, roughly speaking, three kinds of ampliative, non-deductive principles of inference: principles of self-evidence (gnostic access, incorrigibility, a priority, clarity and distinctness, etc.), principles of what might be called 'textbook induction' (enumerative induction, eliminative induction, statistical syllogism, Mill's Methods, etc.), and principles of sophisticated ampliative inference (such as PS principles and the other considerations of theoretical elegance and power mentioned earlier, which are usually construed as filling out the 'best' in 'inference to the best explanation').

P3. Principles of textbook induction are not the interesting principles of rational acceptance in philosophy.

P4. Principles of 'self-evidence,' though popular throughout the history of philosophy and hence considered interesting principles of rational acceptance, cannot be used to settle philosophical disputes.

Therefore:If there are any interesting and decisive principles of rational acceptance in philosophy, they are the elegance principles (p. 119).

Lycan adds that the elegance principles are to be extracted mainly from the history of science, and that we can obtain precise and useful statements of such principles only by looking to the history of science, philosophy, and logic in order to see exactly what considerations motivate the replacement of an old theory by a new theory. (pp. 119-120). So, the answer to the question "Why does it follow that the considerations that make for plausibility in philosophy are the same as those that make for plausibility in science?" is simply "there is nowhere else to turn" (p. 120). For a number of reasons, however, this response to the objection seems problematic.

To begin with, *P*1 is not true. It is common knowledge that one cannot be rationally justified in accepting the conclusion of an argument unless the argument is valid and consistent in addition to the premisses being true. So, we cannot construe *P*1 to assert that validity and consistency are redundant or eliminable as a conditions necessary for the rational acceptability of an argument. Lycan does not mean

to argue that point. Rather *P*1 asserts that even though validity and consistency are necessary conditions for the soundness of an argument, validity and consistency provide no grounds for thinking that the conclusion deduced is plausible. In short, *P*1 asserts is that it is not even a necessary condition for the *plausibility* of a proposed argument that it be both sound and consistent. The reasons offered for *P*1, however, seem particularly questionable and the reasons for thinking *P*1 false seem straightforwardly compelling. Let me explain.

Lycan claims that *P*1 is true because deductive rules are not controversial in their application (p.119). But how exactly would that establish that such rules of inference are not interesting, meaning thereby not plausibility-conferring on the conclusion? Why not say instead that *because* such rules are noncontroversially applied they are interesting, that is, plausibility conferring? In other words, what does the fact that such rules are non-controversial in their application have to do with their not being plausibility-conferring on the conclusions that follow from them? Is it meant to be obvious that a deductive rule of inference is plausibility-conferring only when its application is controversial? Why should anyone accept such a definition of plausibility, especially because it seems to endorse saying such things as "Your argument is perfectly plausible even though it is both invalid and inconsistent." Why aren't deductive rules interesting (or plausibility-conferring) because they are more likely to guarantee truth from premises that are true? If "interest" is relative to purpose and, if one's purpose is to provide a system of inferential rules that is strongly truth-preserving, then such rules are quite interesting, even if their application is non-controversial. This in itself is sufficient to show *P*1 is involved in a questionable bit of semantic legislation.

Moreover, Lycan's second reason for *P*1 is that any deductive argument can be made valid in a perfectly trivial way by the addition of some inference-licensing premise (p. 119). But how exactly does it follow from the fact (if it be a fact) that any deductive argument can be made trivially valid that *no* deductive principle (including consistency) is interesting in the sense of conferring plausibility in any degree on what follows from the deductive principle? Here again, is it meant to be obvious that deductive principles are plausibility-conferring only if they function in arguments incapable of being rendered valid and consistent in non-trivial ways? Does not such a claim presuppose a definition of plausibility which, by stipulation, asserts that the plausibility of a deductive conclusion has nothing to do with the fact that the argument is valid and consistent? And is that not precisely what needs to be shown? Indeed, if any deductive argument could be rendered valid and consistent in wholly trivial ways and the conclusion still be plausible, why insist, as we do, on validity and consistency for soundness as a necessary condition for rational acceptance? Why insist on rules that are truth-preserving for soundness if one can get it in trivial ways and it has nothing to do with the plausibility of the conclusion?

Lycan's third reason for *P*1 is that deductive rules are not plausibility conferring because such rules are uninteresting. They are uninteresting precisely because deductive inferences obviously do not accomplish the expansion of our total store of explicit and implicit knowledge, since they succeed only in drawing out information already implicit in the premises (p. 119). Here again, however, even if deductive inference only renders explicit what is implicitly contained in the premisses that would only show that deductive inference is not inductive inference, and, unless one *assumes* that the only plausibility considerations that will count are those relevant to expanding our factual knowledge base, rather than showing that one's inferences are the product of truth-preserving rules, why would the fact that deductive inference is not inductive inference be a sufficient reason for thinking that deductive inference is uninteresting as a way of enhancing the plausibility of one's conclusions deductively inferred? The reason Lycan offers here (like the two offered above) strongly implies that the plausibility of a person's beliefs has nothing to do with whether it is internally consistent, or follows logically from well-confirmed beliefs, or is consistent with a large body of well-confirmed beliefs, or is the product of truth-preserving rules of inference; and this just flies in the face of our epistemic practices.

Lycan's last reason for *P*1 is that any deductive argument can be turned upon its head (p. 120). Once again, what needs proving is assumed. Even if we can turn a valid deductive argument on its head, so to speak, does that mean that there are no valid arguments? If the answer is yes, why say that valid deductive inference is uninteresting rather than impossible? But if we are not arguing that valid deduc-

tive inference is impossible, why exactly would such inference be uninteresting if it is truth-preserving, and would guarantee consistency, coherence with well-confirmed beliefs, and the explicit addition of true verifiable sentences not formerly in the corpus of our beliefs? If such considerations do not count as plausibility-conferring, it could only be because "plausibility" is stipulatively defined to rule out such considerations as plausibility-conferring. Such a definition needs defending rather than pleading.

By way of general observation with regard to *P*1, it seems clear that plausibility considerations rest quite squarely on questions of consistency and derivability. One of the traditional tests for theory confirmation (and hence by implication for plausibility) is derivability from above. For example, the fact that Balmer's formula for the emission spectra for gases derives logically from Bohr's theory on the hydrogen atom, counts strongly in favor of Balmer's formula above and beyond the evidence Balmer gave for his formula. What is that to say except that considerations purely deductive in nature function to render theories more or less plausible? What about the rest of Lycan's argument against the first objection to his general argument?

Well, suppose, for the sake of discussion that *P*2 and *P*3 are true. Will *P*4 be true? In other words, will it be true that principles of self-evidence do not count for plausibility unless they can be used to settle some philosophical disputes. Here the argument seems to be suggesting that a common-sense principle will be plausibility-conferring only if it can be used to "settle" (in the sense of everybody agreeing henceforth to the answer) some philosophical dispute. But such a requirement seems arbitrarily too strong. Obviously, a conclusion can be plausible and worthy of rational acceptance even when others will disagree to some degree. Two mutually exclusive conclusions may both be rationally plausible without the principle that renders them plausible "settling" the dispute once and for all. Moreover, are we sure that appeals to common sense principles have failed to resolve or settle philosophical disputes? In a very strong sense of "settle," of course, nothing is settled in philosophy. But that would be to impose an arbitrarily strong sense of "settle" on philosophy, a sense we certainly would not impose on science. In a suitably weak sense, "appeals to obviousness or self-evidence" often, but not always, settles disputes. Indeed, isn't the basic reason that the question of solipsism consistently fails to capture anybody's sustained attention is that it is so implausible by way of appeal to common sense? Who these days really takes the possibility of solipsism seriously? Isn't that a philosophical problem pretty much settled by appeal to common sense or self-evidence? Of course, not all appeals to common sense are so successful, and some are more successful than others as clean "conversation stoppers."

These reasons show that Lycan's reply to the first objection fails. Further, it would have been surprising if the reply had succeeded because it seems clear that in science, but not in philosophy, a necessary condition for any explanation being even remotely plausible is that it be in principle empirically testable. As a matter of fact, if we consult practicing scientists and not philosophers, unless one's scientific explanations are ultimately testable, and we know what empirical evidence would need to occur to falsify the hypothesis, we say that the explanation not plausible. We may even go so far as to say that it is meaningless because it is not testable. Minimally, in science a hypothesis or a theory will be plausible only if it is empirically testable, and it will be testable only if what the hypothesis virtually predicts is in principle observable under clearly specifiable conditions, and would occur as expected if we were to accept the hypothesis as worthy. On the other hand, if we are not to beg the question against philosophy as distinct from science, and look at philosophical theses, we will find that a philosophical thesis can be more or less plausible quite independently of whether the thesis is empirically confirmable or testable. As a matter of fact, consider, for example, the dispute between classical scientific realists and classical anti-realists of an instrumental sort. What empirical test might one perform to establish or refute the view that the long-term predictive success of some scientific hypotheses is a function of the truth of claims implied or assumed by the hypotheses? Surely one of these theses must be correct, and yet neither the realist nor the anti-realist position here is testable by appeal to any known experimental or non-experimental test.[9] Does that mean that while one position must be correct *neither* is plausible? Paradox aside, if we say yes, how is that anything more than assuming what needs to be proven, namely that considerations that count for plausibility in science are the same as those that count for plausibility in philosophy?

Surely, however, there are also other philosophical arguments that in fact do depend for their plausibility on the verification and falsification of certain factual claims. For example, Aristotle once argued that humans are quite different from animals because they use tools, whereas animals do not. Aristotle's argument here is implausible because readily falsified by careful observations of the sort Jane Goodall and others continually make. So, in philosophy plausibility may sometimes root in considerations of testability just because one of the premisses in the argument asserts that some factual claim about the world is true or false. But, as we just showed in the case of the dispute between the scientific realist and the scientific anti-realist, plausibility may have very little or nothing to do with the empirical testability of the hypothesis. It may simply be a matter of showing the internal inconsistency of a particular argument, or the dire consequences of adopting one position over the other, or the informal (or formal) fallacies attending the argumentation of one position over the other. In short, as practiced, philosophical reasoning often requires both deductive and inductive principles of rational acceptance for plausibility. But it certainly is not a necessary condition for philosophical plausibility that one's philosophical positions be testable or explainable in a way that accomodates empirical testability as a necessary condition for significance. Moreover, it should be apparent by now that to insist that plausibility in philosophy must acomodate the canons of empirical testability or the canons of explanation in the natural sciences is simply a blatant question-begging move against the objection offered against Lycan's main argument. As such, it would be a rationally unmotivated stipulation against philosophy as distinct from natural science.

In sum, Lycan's reply to the first objection to his general argument fails unless one wants to suppose that there is nothing at all plausible about any philosophical argument primarily because philosophical arguments are not straightforwardly verifiable or falsifiable in the way that empirical claims are. Besides, it seems that the dark shadow of Quine's "Epistemology Naturalized" is having an unrealized effect on the main argument Lycan offers. This is because if one excludes philosophical arguments from the realm of the analytic or a priori, (as Quine does), it would appear that if there is anything to them at all, they must fall into the realm of the synthetic; and hence it seems only too natural to suppose that synthetic claims are meaningful only if testable and confirmable in some basic way by the method of the natural sciences. But the very argument offered from this view supposes, once again, what needs defending, namely, that philosophical plausibility depends on plausibility considerations that are appropriate only to the methods of the natural sciences. When we look to the actual practice of philosophy that assumption seems quite false or the argument Lycan offers begs the question against the distinctness of philosophy. Let's turn to another recent argument for the Quine thesis.

V. The "Traditional Epistemology Will Become Irrelevant" Argument

In his recent book *Explaining Science: A Cognitive Approach.* Ron Giere argues that the justification for the naturalizing of philosophy will not come from explicitly refuting the old paradigm of traditional epistemology, by explicitly refuting on a philosophical basis the philosophical arguments favoring the traditional posture. Rather the argument for naturalizing epistemology will simply be a function of the empirical success of those practitioners in showing how to answer certain questions and, at the same time, showing the irrelevance of the questions asked under the old paradigm.[10] Comparing the naturalised epistemologist with the proponents of seventeenth century physics, he says:

> Proponents of the new physics of the sevententh century won out not because they explicitly refuted the arguments of the scholastics but because the empirical success of their science rendered the scholastic's arguments irrelevant (p. 9).

This same sort of argument has been offered by philosophers such as Patricia Churchland and Paul Churchland, who have claimed that traditional epistemology or "first philosophy" will disappear as a consequence of the inevitable elimination of folk psychology in favor of some future successful neuroscientific account of cognitive functioning.[11]

While there are various reasons for thinking that the eliminative materialism implied by the above argument cannot occur,[12] what seems most obvious is that the assertion made by Giere and the Churchlands is simply not an *argument* for naturalized epistemology. Rather, it is a boyantly optimistic

prediction that, purely and simply because of the expected empirical success of the new model, we will naturally come to regard traditional epistemology (normative epistemology) as having led us nowhere. In short, we will come to view the questions of traditional epistemology as sterile and no longer worth asking. In spite of the optimism of this prediction, it is difficult to see what successes to date justify such a prediction. What central traditional epistemological problems or questions have been rendered trivial or meaningless by the advances in natural science or neuroscience ? Unless one proves that a basic question in traditional epistemology is "How does the Brain Work?" the noncontroversial advances made in neuroscience will be quite irrelevant to answering the questions of traditional epistemology. While some people seem to have *assumed* as much,[13] it is by no means clear that knowing how one's beliefs originate is in any way relevant to their being justified or otherwise worthy of acceptance.[14] Without being able to point to such successes, the eliminative thesis amounts to an unjustified assertion that traditional philosophy is something of a unwholesome disease for which the doing of natural science or neuroscience is the sure cure. In the absence of such demonstrated success, however, no traditional epistemologist need feel compelled by the prediction to adopt the posture of naturalized epistemology.

As a program committed to understanding the mechanisms of belief-acquisition, naturalized epistemology may very well come to show that our traditional ways of understanding human knowledge is in important respects flawed and, as a result, we may indeed need to recast dramatically our understanding of the nature of human knowledge. It would be silly to think that this could not happen. After all, Aristotle's conception of human rationality, and the way in which it was allegedly distinct from animal rationality, was shown to be quite wrong when we all saw Jane Goodall's films showing Gorillas making and using tools. Thereafter, Aristotle's philosophical argument that humans think, whereas animals do not, because the former but not the latter use tools, disappeared from the philosophical landscape. So, it is quite possible that there are certain empirical assumptions about the nature of human knowledge that may well be strongly and empirically falsified in much the same way that Aristotle's position was falsified. But even that sort of progress is still quite consistent with construing epistemology in non-naturalized ways. Traditional epistemology should have no difficulty with accepting the view that some philosophical theses can be conclusively refuted by the occurrence of certain facts. That would be simply to acknowledge that philosophy, and philosophical arguments, are not purely a priori and hence immune from rejection by appeal to the way the world is. So, the traditional epistemologist will need to wait and see just what naturalized epistemology comes up with. Whether it lives up to the expectations of Giere and others who, like the Churchlands, offer the same basic argument is still an open question, at best. As things presently stand, there are good reasons, as we shall see, for thinking that no amount of naturalized epistemology will ever be able in principle to answer certain crucial questions about the nature of justification.

Otherwise Giere's defense of naturalized epistemology consists in responding to others who argue against naturalized epistemology. In responding to these objections, Giere seeks to show that there is certainly no compelling reason why one should not proceed on the new model. He considers the following three arguments.

A. *Putnam's Objection*

In his "Why Reasons Can't be Naturalized" (*Synthese*, vol. 52, 1982, pp. 3-23) Putnam says:

> A cognitive theory of science would require a definition of rationality of the form: A belief is rational if and only if it is acquired by employing some specified cognitive capacities. But any such formula is either obviously mistaken or vacuous, depending on how one restricts the range of beliefs to which the definition applies. If the definition is meant to cover *all* beliefs, then it is obviously mistaken because people do sometimes acquire irrational beliefs using the same cognitive capacities as everyone else. But restricting the definition to rational beliefs renders the definition vacuous. And so the program of constructing a naturalistic philosophy of science goes nowhere (*Synthese*, vol. 52, 1982, pp. 3-23).

In response to this particular argument, Giere says:

> The obvious reply is that a naturalistic theory of science need not require any such definition. A naturalist in epistemology, however, is free to deny that such a conception can be given any coherent content. For such

> a naturalist, there is only hypothetical rationality which many naturalists, including me, would prefer to describe simply as 'effective goal-directed action,' thereby dropping the word 'rationality' altogether (p. 9).

In short, for Giere, Putnam is just begging the question by insisting that there must be a coherent concept of categorical rationality. In defense of Putnam's intuition, however, one can argue that Giere missed Putnam's point. Putnam's point is just as easily construed as asserting that if the naturalised epistemologist is not to abandon altogether the concept of rationality, (and thereby abandon any way of sorting justifiable or warranted beliefs from those that are not) the rationality of a belief will be purely and simply a function of the reliability of the mechanisms that cause the beliefs. But because a belief can be produced by a reliable belief-making mechanisms and be rationally unjustified, such a definition will not work. Putnam's objection, when construed in this way, is compelling. Unfortunately, Giere's response seems to miss the point Putnam makes. Presumably, Putnam would respond that even if we were to stop talking about rationality, we would still need some way of determining which beliefs are more or less justified; and the naturalized epistemologist would need to define such concepts in terms of the mechanisms that produce certain beliefs. And Putnam's point is that that just will not work because unjustified beliefs can emerge just as easily from reliable mechanisms.

B. *Siegel's Objection*

In his "Justification, Discovery and the Naturalizing of Epistemology,"[15] Harvey Siegel challenges the naturalistic approach by arguing that rationality of means is not enough. There must be a rationality of goals as well because there is no such thing as rational action in persuit of an irrational goal. In response to this objection, Giere notes:

> This sort of argument gains its plausibility mainly from the way philosophers use the vocabulary of 'rationality.' If one simply drops this vocabulary, the point vanishes. Obviously, there can be effective action in persuit of any goal whatsoever-as illustrated by the proverbial case of the efficient Nazi"....
>
> Nor does the restriction to instrumental rationality prevent the study of science from yielding normative claims about how science should be pursued. Indeed, it may be argued that the naturalistic study of science provides the only legitimate basis for sound science policy (Campbell, 1985). (p. 10.)

Along with Giere, we may find it difficult to take Siegel's objection seriously for two reasons. Firstly, it is not at all obvious that naturalized epistemology is committed to rationality of means only and not also to the rationality of goals. It is not even clear what that claim amounts to. Secondly, as Giere also points out, there certainly seems to be cases in which irrational ends can be pursued by rational action. Anyway, as we shall see later, there are much more persuasive objections to naturalized epistemology.

C. *The Objection from Vicious Circularity*

There is another common objection to eliminating traditional epistemological questions in favor of questions about effective means to desired goals. Giere characterises it in the following way:

> To show that some methods are effective, one must be able to show that they can result in reaching the goal. And this requires being able to say what it is like to reach the goal. But the goal in science is usually taken to be 'true' or 'correct' theories. And the traditional epistemological problem has always been to justify the claim that one has in fact found a correct theory. Any naturalistic theory of science that appeals only to effective means to the goal of discovering correct theories must beg this question. Thus a naturalistic philosophy of science can be supported only by a circular argument that assumes some means to the goal are in fact effective." (p. 11.)

Giere then proceeds to show that this sort of objection (which he does not cite anybody in fact offering) is based on some dubious items of Cartesian epistemology. A more direct response, however, is that this objection is unacceptable because it assumes rather than proves that the goal of scientific theories is to achieve truth rather than empirical adequacy. In other words, a proper response would consist in straightforwardly denying that the goal of science is to discover "true" or "correct" theories rather than ones that are instrumentally reliable as predictive devices. So, for other reasons, we need not take this objection very seriously. In the end, apart from Putnam's objection, these last two objections to naturalized epistemology do not have the necessary

bite, and Giere seems quite justified in rejecting them. Later we shall see better objections. For now, however, we need only note that the above argument in favor of naturalized epistemology is not an argument, and that the author's response to the above three objections to naturalized epistemology selects only three and fails to deal effectively with Putnam's objection.

In the end, for Giere, evolutionary theory provides an alternative foundation for the study of science:

> It explains why the traditional projects of epistemology, whether in their Cartesian, Humean, or Kantian form, were misguided. And it shows why we should not fear the charge of circularity (p. 12).

But what exactly is it about traditional epistemology that made it misguided? That it sought to refute universal scepticism? Whoever said that that was *the goal* of traditional epistemology? As we noted earlier when we examined Quine's argument, to define the concept of knowledge and then to determine whether, and to what extent, human knowledge exists in the various ways we define it seems equally the major goal of traditional epistemology. And why, exactly, is that a misguided activity? Such an activity seems justified by the plausible goal that if we get very clear on just what we mean by basic epistemological concepts, we might just be in a better position to determine the snake-oil artist from those whose views are worthy of adoption. This goal is based on the noncontroversial point that knowledge just isn't a matter of accepting everything a passerby might say. At the root of most arguments for naturalized epistemology, as we shall see, is this peculiar claim to the effect that traditional epistemology somehow has failed or been misguided in its search for some cosmic skyhook. Certainly we saw as much when we examined Quine's argument. But when the arguments are laid on the table, some philosophers may come to think that what gets characterized as tradtional epistemology is quite different from the real thing. Socrates, after all, began his discussion in *Theatetus* with the question "What is knowledge?" and not "Is Human knowledge possible?" or "How does the mind represent reality?" or (as one philosopher recently claimed) "How does the brain work?" [16] That anybody could seriously think that Socrates was really asking for an account of how the brain works is difficult to comprehend. And to say that that is what he *should* have been asking (because nobody has or can answer whatever other question he might have asked) presupposes that one can show that the questions he did ask are misguided or bad questions, and that is yet to be shown in any way that does not beg the question against philosophy. We may now turn to the fourth argument in favor of Quine's thesis.

VI. The Argument From Evolutionary Theory

Evolutionary epistemology is a form of naturalized epistemology which insists that the only valid questions about the nature of human knowledge are those that can be answered in biological science by appeal to evolutionary theory. For the evolutionary epistemologist, the Darwinian revolution underscored the point that human beings, as products of evolutionary development, are natural beings whose capacities for knowledge and belief can be understood by appeal to the basic laws of biology under evolutionary theory. As Michael Bradie has recently noted, evolutionary epistemologists often seem to be claiming that Darwin or, more generally, biological considerations are relevant in deciding in favor of a non-justificational or purely descriptive approach to the theory of knowledge.[17] When we examine the arguments proposed by specific evolutionary epistemologists, there seems to emerge two distinct arguments. The first argument, allegedly offered by philosophers such as Karl Popper, and reconstructed by Peter Munz is as follows:

> *P*1. We do in fact have human knowledge.
>
> *P*2. No justification is possible.
>
> Therefore: *P*3. Human knowledge does not involve justification.
>
> Therefore: *P*4. Every item of knowledge is a provisional proposal or hypothesis subject to revision.[18]

For this reason Popper held that the only problem in epistemology was the problem of the growth of human knowledge, or the biological question of how human knowledge originates and grows. Therefore, epistemology is not normative in the way that traditional epistemology is normative but rather purely descriptive. As Bradie has noted, Popper's argument for *P*2 is based on his acceptance of Hume's critique of induction and the corrolary that no empirical universal statements are provable beyond doubt. (p. 10.) The second argument, inspired by Quine's ref-

erence to Darwin, is offered by Hilary Kornblith and reconstructed by Bradie as follows:

> *P*1. Believing truths has survival value.
>
> Therefore: *P*2. Natural selection guarantees that our innate intellectual endowment gives us a predisposition for believing truths.
>
> Therefore: *P*3. Knowledge is a necessary by-product of natural selection

In order to get the desired conclusion of a purely descriptive epistemology, Kornblith supplies the following premise

> Therefore: *P*4. If nature has so constricted us that our belief-generating processes are inevitably biased in favor of true beliefs, then it must be that the processes by which we arrive at beliefs just are those by which we ought to arrive at them.

warranting the final conclusion:

> Therefore: *P*5. The processes by which we arrive at our beliefs are just those by which we ought to arrive at them.[19]

What can we say about these two arguments?

With regard to the first argument, the one Peter Munz ascribes to Popper, the first thing to note is that there is nothing particularly "biological" or "evolutionary " about it at all. It is simply a philosophical argument based on a philosophical acceptance of Hume's philosophical scepticism to the effect that no factual claim about the world could be justified sufficiently for knowledge. So, the argument does not provide a justification deriving from evolutionary theory for taking the naturalistic turn. Secondly, as we saw when we discussed Quine's argument above, accepting a philosophical argument for a purely descriptive epistemology is radically incoherent. A philosophical argument to the effect that there is no first philosophy because Hume was correct in his defense of the problem of induction, is radically incoherent and self-defeating in a way apparently not yet appreciated by naturalized epistemologists.

The second argument has already been well-criticised by Michael Bradie who has noted (along with many others,including Stitch, Leowontin, and Wilson) that *P*2 is quite questionable. The fact that certain beliefs endure and have survival value by no means implies that they are the product of natural selection. There are many traits that evolve culturally which have no survival value (See Bradie, p. 16) Moreover, even if it were true that our cognitive capacities have evolved by natural selection, the important point is that that by itself is no reason for thinking that we are naturally disposed to believe truths rather than falsity. On the contrary, the evidence seems pretty strong that, given the history of scientific theorizing, the species is more disposed to accept empirically adequate rather than true hypotheses.

One interesting response to this last line of reasoning comes from Nicholas Rescher who has argued in *Methodological Pragmatism* that say what we will, the methods of the natural sciences have indeed been selected out by nature, otherwise they would not have endured as such reliable instruments for prediction and control. Rescher's basic point is that on any given occasion, an instrumentally reliable belief or thesis may well fail to be true. But that is no reason for thinking that nature has not selected out the methods of the natural sciences because in the long run the methods of the natural sciences provide truth.[20] Rescher's point is well taken, but it is certainly not an argument for the thesis that epistemology is purely descriptive. Rescher certainly is not a naturalized epistemologist in that sense. Rather it is an argument for regarding the deliverances of the methods of natural science as epistemically privileged. In offering the argument he does here, Rescher is merely showing how the usual arguments against Pragmatism hold for *thesis* pragmatism and not for *methodological* Pragmatism. Nor does his argument provide the evidence necessary for making sound Kornblith's reconstructed argument from evolution. This is because Kornblith's proposed argument still falters on *P*2. Rescher's argument by no means shows or supports the view that people by nature are innately disposed to believe only true propositions. If that were so, it would be difficult to see why we would ever need the methods of the natural sciences anyway. Nature selected out the methods of the natural sciences just *because* we are not natively disposed to believe only true propositions.

But, if the above two arguments are the best evolutionary biologists can offer in defense of the first form of naturalized epistemology, it would seem that biology itself, and especially evolutionary biology, is yet to offer a persuasive argument for naturalized epistemology. Along with Bradie, we can only conclude that there does not seem to be any

persuasive argument from evolutionary theory in favor of the first form of naturalized epistemology.[21]

VII. The "Impossibility of Defining Justification" Argument

The last, and perhaps the most interesting, argument for the first form of naturalized epistemology appears in a forthcoming paper by Richard Ketchum entitled "The Paradox of Epistemology: A Defense of Naturalism." Ketchum's argument is the following: An adequate traditional epistemology will require, among other things, an acceptable definition, or explication of the concept of justification. But there is no non-question begging definition, or explication of the concept of justification. This latter claim rests on the reason that whatever definition one would offer for the concept of justification admits of the question "Are you justified in accepting or believing this definition of justification?" And, of course, if one were to answer yes and then defend the answer by saying it is an instance of the definition (which presumably one would need to say), the questioner would reply that the appeal is question-begging because what is at issue is whether *that* definition itself is justified. Appealing to the analysans of the definition to justify the definition is a patent bit of question-begging. So, no matter what one's definition might be, there would be no non-question begging way of answering the question of whether one is justified in accepting that definition. Thus, traditional epistemology is dead.[22] What about this argument?

One possible response is that while one may not be justified in believing one's definition of justification, one might certainly have good reasons for accepting one's definition of justification. But the problem with this response is that it arbitrarily prevents one from defining justification in terms of having good reasons. Besides, if having good reasons for accepting a definition of justification is sufficient for accepting it, then why is that not the definition of justification? Can one have sufficient reasons for accepting something and not be justified in accepting it?

Another possible response asserts that the problem with this argument is not in the assumption that we must be justified in believing our definition of justification. Rather it is in the assumption that justification in believing a definition has the same meaning as justification when the term applies to non-definitions. On this view, being justified in believing a definition is simply a matter of whether one has correctly generalised from the conditions of correct usage in natural or scientific discourse (or, if our definitions are stipulative, a matter of whether they lead us to conclusions that satisfy the purpose behind defining things the way we do); whereas being justified in believing a non-definition, or a proposition about the world, is a matter of whether one can give (if necessary) good reasons for thinking that the proposition is a reasonably adequate description of one's mental content or of the non-mental world. Is there anything wrong with this proposed solution?

Yes, and it is this: The original question returns in the form of the question "Are you justified in believing that the concept of justification differs for reportive definitions and non-definitions in the way indicated?" Here again, if one answers affirmatively, one could only defend the answer by making it an instance of the concept of justification appropriate for non-definitions, and that is what is at issue. In short, the question returns with a sting even when we try to distinguish various senses of justification.

By implication, suppose one were to say "I am justified in accepting my definition of justification because the definition conforms to the rules we require for generating acceptable definitions." Once again, the obvious response is "Are you justified in accepting the rules for generating acceptable definitions?"; if the answer is yes (as presumably it would be), then the answer is defensible only if it is an instance of one's definition of justification for non-definitions; but one's definition of justification for non-definitions is justifiable only if it is an instance of the definition of justification for non-definitions. But the latter is itself what is at issue, and so we come back to the original question and the impossibility of answering it in a non-question-begging way.

Yet another response consists in trying to rule against the meaningfulness of the question on the grounds that if we take it seriously, then it would lead to an infinite regress and that in itself is good evidence for the inappropriateness of the question. On this view, whatever answer one gives, the respondent could still ask "But are you justified in believing that?" Differently stated, to countenance the question in the first instance is to countenance more properly the assumption that one must be justified in all one's beliefs and, as we know from

Aristotle's argument in the first book of *Prior Analytics*, that requirement guarantees scepticism, because the need for an infinite amount of justification prevents there ever being any demonstrative knowledge. Is this response acceptable?

It is not acceptable as a way of establishing non-traditional epistemology because the same question cuts equally strongly against the naturalized epistemologist. The naturalized epistemologist, like Quine, still says that one's beliefs about the world are more or less justified by appeal to the canons of scientific inference. Accordingly, suppose we grant that traditional epistemology is dead and that one could still be justified in one's beliefs about the world because we need only follow the canons of justification as practiced in science. But if the question "Are you justified in believing that your definition of justification is appropriate or correct?" is a legitimate question to ask of the classical epistemologist, it is also a legitimate question to ask of the naturalized epistemologist who asserts that "In natural science, being justified in one's beliefs is simply a matter of x." And if this is so, the nature of justification in science is as problematic as it would be in traditional epistemology. Why is it that for the naturalized epistemologist the natural scientist (but not the traditional epistemologist) can well ignore the philosopher's question " Are you justified in accepting x as the correct definition of justification in science?" Indeed, it would seem that the question cuts both ways, and is not any more devastating for the traditional epistemologist than it is for the practice of science in general or for the naturalized epistemologist. If the naturalized epistemologist feels justified in ignoring such questions because they are so obviously philosophical, why exactly is that an argument in favor of naturalized epistemology rather than an unargued rejection of philosophy in general? Thus if the question is persuasive, it tends to show the truth of skepticism in general and not simply the failure of traditional epistemology. This is hardly a desirable result for anybody.

By way of confronting this pesky argument for the impossibility of traditional epistemology, another interesting response consists in asserting that we must begin by accepting the fact that we know something, and that just means that we must reject any and all questions about human knowledge that can only be answered with a question-begging response. So, the truth of the matter is that there is no non-question begging way to answer questions such as "Are you justified in believing your definition of justification?" But if we insist on answering such questions we make global skepticism certain. Presumably, even naturalized epistemologists do not want to go that far. Consequently, such questions are not permissible. What this means is that generalizing from the facts of ordinary usage and scientific practice to determine what we mean by certain epistemic concepts is simply where we start and what we do to get clear about what human knowledge is. To ask that we be justified in the conclusions we draw here is to demand that we begin somewhere else when there is in fact nowhere else to go, and if we do not stop here or somewhere else (which will certainly happen if we allow the skeptic's eternal question " But are you justified in believing that?") there could be no knowledge about anything at all.

Is this a compelling response, or is it merely a grand way of begging the question against global skepticism which, if the original question is permissible, turns out to be forceful? Are we dismissing the question as meaningful because otherwise we would need to accept global skepticism? It is tempting to think not, but the skeptic will doubtless see things differently. It appears that we have no knock-down argument against the skeptic except to say that he begins with a view that we cannot accept, namely, that it is possible that nobody knows anything. But that is not an argument.

In the end, however, the best way to confront the claim that there is no non-question begging way to justify any definition or analysis of justification consists in arguing as follows: Whoever asks the question "Are you justified in accepting your definition of justification?" can be met with the response "What do you mean when you ask whether I am justified in accepting this definition?" When anyone asks the question "Are you justified in accepting or believing your definition of justification?" he must have in mind just what it means to be justified, otherwise it is not a meaningful question because if he did not have in mind just what it meant, he would not know what would count for a good answer if an answer were possible. So, if the question makes any sense at all, the questioner must be prepared to say just what he means when he asks the question "Are you justified in believing your definition of justification?" In fact, then, it is a necessary condition for this question being meaningful that the questioner be able

to say what it would mean for someone to be justified in believing that a particular definition of justification is correct. If the questioner cannot answer the question "What do you mean?" then the question need not be taken seriously. If he can, then his question is easily answered. For example, if the questioner is asking for a good reason for accepting the definition, the response might well be that we have a good reason because the definition is a sound generalisation of the facts of ordinary usage (and that's a good reason because evolution selects out this way of determining the meaning of expressions).

In sum, if the question is an honest one, then the questioner is asking for a justification and if he cannot say what would count as an answer to his question (thereby saying what he means by justification") then we need not, and will not, take his question seriously. On the other hand, as soon as he tells us just what he means by "justification" his question seems meaningful and answerable. But now comes the rub: we can still refuse to take his question seriously because *we* can now raise the question of whether his understanding of "justification is justified, because if it is not, we do not need to answer his question; and if he says our question is meaningless, then so too was his initial question. But now the shoe is on the other foot, as it were. The person who questions the original definition of justification can make sense of his question only if he is willing to say just what justification consists in; but if the original question makes sense then it will make equal sense when the question is asked of him—meaning that *he* has no non-question begging way of answering a question necessary for his meaningfully asking "Are you justified in believing your definition of justification?" Thus it appears that we are justified in ignoring the question because the questioner cannot, ex hypothesi, satisfy a condition necessary for the meaningfulness of the proposition. He cannot, ex hypothesi, answer in any non-question begging way the question we can ask of him, namely. " Are you justified in accepting your definition of justification?"

VIII. Conclusion

Given the above considerations, it seems that, in spite of the popularity of the thesis, there is no sound argument presently available supporting the Quinean version of naturalized epistemology. Nor should we be tempted to suppose that because we have never achieved a consensus in traditonal epistemology, it looks as though we have good inductive grounds that the program of traditional epistemology will never work. That sort of argument blatantly begs the question in favor of a concept of success that is appropriate to the methods of the natural sciences and so, by implication, begs the question in favor of the naturalized epistemology for which it is supposed to be an argument. The interesting question is whether there is something fundamentally incoherent about arguing philosophically for such a naturalized epistemology. As was suggested above in the discussion on Quine's argument, it certainly seems that offering a philosophical argument in favor of denying that philosophical arguments will count when it comes to answering questions about the nature of epistemology is incoherent when the point of it is to defend a particular view about the nature of human knowledge. But perhaps this is merely a philosophical point.[23]

Georgia State University

Received October 13, 1989

NOTES

1. In *Ontological Relativity and Other Essays,* New York: Columbia University Press, 1969).

2. For a full discussion of Goldman's thesis as it occurs in *Epistemology and Cognition* (Cambridge, Mass: Harvard University Press, 1985) see R. Almeder and F. Hogg, "Reliabilism and Goldman's Theory of Justification," *Philosophia* (vol. # 1989) pp.

3. In Hilary Kornblith, ed., *Naturalizing Epistemology* (Cambridge: MIT Press, 1985), pp. 71-85.

4. *Roots of Reference* (LaSalle: Open Court, 1975).

5. In S. Guttenplan, ed., *Mind and Language* (Oxford: Clarendon Press, 1975).

6. See "Reply to Stroud" in *Midwest Studies in Philosophy Vol.VI,* P. French, E. Uehling, and H. Wettstein, (eds) (Minneapolis: University of Minnesota Press, 1981), p. 474.

7. See "Nature Unmirrored, Epistemology Naturalized" in *Synthese* vol. 55 (1983), p. 69.

8. See *Judgement and Justification* (Cambridge: Cambridge University Press, 1988).

9. This same point is made by P. Skagestadt in "Hypothetical Realism" in Brewer and Collins (eds.) *Scientific Inquiry and the Social Sciences: A Volume in Honor of Donald T. Campbell* (San Francisco: Jossey-Bass, 1981), p.92.

10. Ron Giere, *Explaining Science: A Cognitive Approach* (Chicago: University of Chicago Press, 1988).

11. See P.S. Churchland's *Neurophilosophy* (Cambridge: MIT Press, 1986); P. M. Churchland's "Eliminative Materialism and the Propositional Attitudes," *Journal of Philosophy* (1981), vol. 78, pp. 67-90 and "Some Reductive Strategies in Cognitive Neurobiology, *MIND* (1986), vol. 95, pp. 279-309. For a similar argument see S. Stitch's, *From Folk Psychology to Cognitive Science* (Cambridge: MIT Press, 1983).

12. For an interesting argument to the effect that folk psychology is not likely to be eliminated in the way the Churchland's assert that it will, see Robert McCauley's "Epistemology in an Age of Cognitive Science," *Philosophical Psychology*, vol. 1 (1988), pp.147-149.

13. See, for example, P. S. Churchland's "Epistemology in an Age of Neuroscience," *The Journal of Philosophy*, vol. LXXXIV, No. 10 (Ocy. 1987), p. 546 where she asserts, without benefit of argument, that the basic question in epistemology is indeed the question "How does the brain work?" Also, as an interesting example of an argument seeking to show that traditional epistemological questions can be eliminated owing to advances in cognitive science (advances that presumably show how we can understand human knowledge on a non-propositional basis), see William Bechtel and Adel Abrahamson, "Beyond the Exclusively Propositional Era," *Synthese*, vol. 82 (1990).

14. This same point is made by Ernest Sosa in "Nature Unmirrored: Epistemology Naturalised," *Synthese*, vol. 55 (1983), p. 70. Naturally, if one were able to defend a form of reliabilism similar to that offered by Alvin Goldman, the question "How does the brain work?" would turn out to be a most crucial question in epistemology; but at that point we would not be defending the sort of naturalised epistemology defended by Quine rather than the form offered and defended by Goldman.

15. See Harvey Siegel, *Philosophy of Science* vol. 47 (1980), pp. 297-321.

16. See P.S. Churchland, "Epistemology in the Age of Neuroscience," *Journal of Philosophy* (Oct. 1987), p. 546.

17. See Michael Bradie, "Evolutionary Epistemology as Naturalised Epistemology," (forthcoming), p. 3.

18. See Peter Munz, *Our Knowledge of the Growth of Knowledge: Popper or Wittgenstein*? (London: Routledge and Kegan Paul, 1987), p. 371). For a defense of a similar argument, see also Michael Ruse's *Taking Darwin Seriously: A Naturalistic Approach to Philosophy*, (Oxford: Blackwell, 1986), and W.W. Bartley III, "Philosophy of Biology versus Philosophy of Physics," in G. Radnitzky and W. W. Bartley, III (eds.) *Evolutionary Epistemology, Theory of Rationality and the Sociology of Knowledge*, LaSalle: Open Court, 1987), p. 206).

19. See Hilary Kornblith, *Naturalizing Epistemology* (Cambridge: MIT Press, 1985), p. 4.

20. Nicholas Rescher, *Methodological Pragmatism* (Oxford: Basil Blackwell, 1977), Ch. 6

21. For further reasons why evolutionary epistemology has in fact failed to offer any compelling explanation of how what we take to be knowledge, especially scientific knowledge, has developed, see William Bechtel's "Toward Making Evolutionary Epistemology Into a Truly Naturalized Epistemology" in N. Rescher, *Evolution, Cognition and Realism* (Washington, D.C.: University Press of America, 1990).

22. This is an informal reconstruction of the argument defended by Richard Ketchum in his forthcoming paper in *Philosophical Studies.*

23. I would like to thank the Center for the Philosophy of Science at the University of Pittsburgh for the Senior Fellowship and the stimulating environment that made the writing of this essay possible. For the same reason, I am grateful to the Hambridge Center. Also, Nicholas Rescher, David Blumfeld, Bill Bechtel, Milton Snoeyenbos, Douglas Winblad, and Richard Ketchum all provided valuable comments and criticisms.

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 27, Number 4, October 1990

RECENT WORK ON NATURALIZED EPISTEMOLOGY

James Maffie

A growing number of philosophers call for our naturalizing epistemology.[1] Yet understanding what this implies is obscured by the confounding variety of its proponents. Philosophers as diverse as Boyd, Dewey, Goldman, Hooker, Laudan, Quine, and Putnam (to name just a few) all hoist the banner of naturalism. Does a single naturalist theme unite such unlikely bedfellows, or are there simply as many naturalisms as there are naturalists? How do naturalists differ from non-naturalists? And finally, how do naturalized epistemologies differ from one another?

Continuity between epistemology and science lies at the heart of these puzzles. Naturalists are united by a shared commitment to the continuity of epistemology and science. Naturalist and non-naturalist divide over whether or not continuity exists. Naturalists differ among themselves over what form this continuity should take. This essay surveys recent work on naturalized epistemology from the perspective of continuity. Part I sketches some background and motivation for naturalizing epistemology. Part II distinguishes six kinds of continuity defended by naturalists. Part III examines debate among naturalists regarding the metaphysical and analytic continuity binding science and epistemology. Part IV considers debate over the proper scope of epistemology's naturalization. I explore some difficulties facing criterial forms of naturalism in Part V.

I. BACKGROUND AND MOTIVATION

Debates concerning the possibility of having knowledge of the external world have been largely influenced by the following intuitive picture:

(1) epistemology should provide a grounding for our belief concerning the external world;

(2) the subject of experience and the object experienced are distinct: the external world exists independently of the sensations, theories, or evidential practices of the subject;

(3) the subject lacks direct or immediate access to the external world and hence needs internally available criteria of knowledge or justification;

(4) truth consists of a non-epistemic relationship (such as correspondence) between belief and world.

Together, these seem to lead irresistibly to skepticism. Attempts to construct a presuppositionless proof of the epistemic credentials of our beliefs concerning the external world have met with little, if any, success. Traditional internalist strategies such as foundationalism and coherentism no longer hold any promise. As a result, many acquiesce to some form of skepticism.

Others, however, find skepticism unacceptable and turn instead to diagnosis. Re-examining the picture above, they hope to isolate and then revise the element(s) responsible for this outcome. Not surprisingly, different elements are identified. Some diagnose skepticism as arising from the "Platonic" or realist conceptions of existence and truth expressed in (2) and (4), and consequently prescribe our reconceiving truth and existence in normative terms. Others point to the specifically *epistemological* elements of the picture and the conception of inquiry they suggest. Skepticism arises from the demand for internally accessible criteria of knowledge (element 3) in conjunction with the demand for an epistemological grounding for our beliefs concerning the external world (element 1). Together, these yield a conception of epistemology according to which the successful grounding of our beliefs can only be achieved via an internally available, apriori self-evident proof. Epistemology is to be conducted as "First Philosophy."

Naturalists challenge this picture of epistemic inquiry on grounds ranging from its fruitlessness to its

unintelligibility.[2] Hooker (1987) and Suchting (1983) reject first-philosophy-style epistemology on grounds of its internal inadequacy. It inevitably resorts to dogmatism, i.e. to making claims that are incapable of being justified by the theory of knowledge subsequently developed. For example, empiricism claims sense experience to be the source of knowledge—but the claim cannot be known via sense experience. The project of first philosophy is hoisted on its own petard since it ineluctably makes assumptions that are illegitmate by its own standards. According to Lycan (1988) the demand that we start from "an epistemic position of zero" is contradictory, in principle impossible to fulfill, and hence unworthy of further effort. The "ought" of first-philosophy-style epistemology no longer binds us since we cannot be obliged to perform contradictory tasks; it violates the principle of "ought implies "can." Goldman (1976), Boyd (1984), Devitt (1984) and Quine (1974) consider the skeptic's challenge intelligible yet unanswerable because its standards are too high. They concede our inability to rebutt skepticism but no longer care since the project is hopeless and no longer interesting. Giere (1985) argues the skeptic presupposes an outmoded and mistaken theory of knowledge-acquisition and thus need no longer be worried about. Kornblith (1988) contends the skeptic holds an empirically false theory about the comparatively greater epistemic access we have to our mind as opposed to the external world; while Hooker (1987, p. 262) characterizes the skeptic as "moronic" about cognition. Lastly, Putnam (1981) denies the intelligibility of the skeptical thesis that our best theories of the world may turn out to be false.

Naturalists maintain it is therefore rational to abandon first-philosophy-style epistemology and pursue instead an alternative with some hope of success. The "cure" for our skeptical ills involves a radical reworking of our idea of epistemology itself. Epistemology must undertake a "naturalistic turn." Dewey (1948, 1958) suggests that by extending the experimental methods of scientific belief regulation to epistemology we make possible an improvement in philosophical understanding analogous to that already effected in the study of physical nature. Armstrong (1978, p. 274) advocates adopting natural methods "since the best guide we have to the nature of reality is provided by natural science." Piaget urges epistemology to follow the example of physics, biology, etc. and break with armchair speculation.[3] In short, naturalists set out to "reconstruct" (Dewey, 1948) epistemology so as to build a "post-Cartesian" epistemology (Goldman, 1978).

II. The Continuity of Epistemology and Science

Naturalists reject the autonomy of epistemology, seeking to create *continuity* between epistemology and natural science. They seek epistemological, contextual, and methodological continuity between the two. Naturalized epistemology employs the cognitive methods of science, adopts the substantive claims of science, and enjoys the aposteriori evidential status of science. Naturalizing epistemology also involves creating analytic, metaphysical, and sometimes axiological continuity between epistemology and science. I consider each in turn.[4]

(A) *Contextual continuity*: Naturalists maintain that epistemology takes place within the context of natural science. It occurs as an enterprise within science, neither coming before nor standing outside of science. Epistemologist and scientist alike stand aboard Neurath's boat as they cooperate in constructing our overall account of the world. Naturalists thus reject non-contextualism or the idea that epistemology must be conducted without presupposition, from an archimedean vantage point. "There is no point of cosmic exile" (Quine, 1960, p. 275). We begin, epistemically speaking, with the beliefs we have at any given time.

> In studying and criticizing our cognitive procedures, we should use whatever powers and procedures we antecedently have and accept. There is no starting 'from scratch'..."[5]

Epistemology makes free use of the findings in neurophysiology, cognitive psychology, evolutionary biology, etc. In fact, if it is to construct the best possible theory of knowledge, it is incumbent upon it to make use of the best theories available.

Naturalism forsakes the foundational task of providing an absolute validation of our common sense or scientific beliefs. Instead, it seeks "to derive the reliability of our methods from our psychological and physical theories, our theories about how the human mind arrives at beliefs through interaction with its environment" (Friedman, 1979, p. 370). It tries to establish the reliability of our methods not in

every possible world but in the actual world as specified by science. The warrant-conferring capacity of our cognitive processes is a contingent fact about the *actual* world.

> I take [naturalized epistemology] to mean that we have in general no apriori way of knowing which strategies for forming and refining our beliefs takes us closer to the truth. The only way we have of proceeding is to assume the approximate truth of what seems to us the best overall theory we already have of what we are like and what the world is like, and to decide in the light of *that* what strategies of research and reasoning are likely to be reliable in producing a more nearly true overall theory. (Sturgeon, 1984, p. 67)

Naturalism thus seeks solutions to our epistemic puzzles within the defeasible substantive context of the assumed truth of our best scientific theories about the nature of human cognizers, their environment, etc. It asks, "In a world of such specifications, is it possible for cognizers with such characteristics to acquire knowledge? What sorts of cognitive strategies are likely to be reliable?" Giere (1985, p. 339) writes:

> "The general problem faced by a naturalistic philosophy of science ...is to explain how creatures with our natural endowments manage to learn so much about the detailed structure of the world—about atoms, stars and nebulae, entropy, and genes. This problem calls for a *scientific* explanation."

Skepticism itself is naturalized, recast as an empirical problem to be addressed by scientific epistemology operating from the perspective of our best current scientific theories. "Skeptical doubts are scientific doubts," (Quine, 1975a, p. 68). Naturalists typically leave open the possibility of a scientifically-based skepticism: whether or not we have knowledge of the external world is a contingent matter to be decided on the basis of the picture of human cognitive processes, etc., given to us by science.

(B) *Epistemological continuity*: Epistemology, like science, is an aposteriori enterprise. Epistemic issues are empirical and resolvable a posteriori. Evaluating the epistemic status of belief, for example, appeals to a posteriori evidence as "it is an empirical question which procedures are good," (Devitt, 1984, p. 66). Epistemic questions are "scientific questions about a species of primates, and they are open to scientific investigation," (Quine, 1975a, p. 68). Naturalized epistemology thus moves away from claims of special evidence, be it divine, apriori, scriptural, or the "dictates of pure reason."

Skepticism, for example, becomes an a posteriori issue. Whether or not we have knowledge of the external world is decided on a posteriori grounds. The warrant-conferring capacity of our cognitive processes is a contingent fact about the world and as such decidable aposteriori. Anti-realists like Laudan and Van Fraassen disagree with scientific realists like Boyd, Hooker and Friedman over whether or not there are good aposteriori reasons for thinking the theoretical assertions of science are in fact warranted. Devitt (1984) and Boyd (1984) disagree with Van Fraassen (1980) and Fine (1984) over the a posteriori status of abductive inference.

Van Fraassen (1980, p. 57) reconstructs the observable vs. non-observable dispute in a posteriori terms: "To find the limits of what is observable in the world described by a theory T we must inquire into T itself, and the theories used as auxiliaries in the testing and application of T." Similarly, Giere (1985) recasts the context of justification vs. context of discovery distinction as an a posteriori distinction representing the hardwon fruit of practical experience. It constitutes part of our empirical theory of science and the world and as such is subject to empirical revision.

Piaget (1971, 1972), Hooker (1974), Haack (1975), and others argue epistemological views are subject to empirical refutation since they presuppose *factual* claims. For example, a posteriori evidence showing the unreliability or non-existence of special cognitive faculties for apprehending conceptual or necessary truths undermines many philosophical accounts of a priori knowledge; that showing the unreliability or non-existence of special faculties for apprehending non-natural, normative properties undermines non-naturalist views in epistemology and ethics; and that showing the theory and value ladenness of sense perception undermines basic tenets of Humean empiricism.[6]

(C) *Methodological continuity*: Naturalized epistemology employs ordinary scientific methods (e.g., observation, testing, induction, experimentation) in resolving problems. Epistemology becomes "science self-applied," (Quine, 1975b, p. 293). It thus moves away from claims of special, distinctly philosophical methods.

Proper philosophical method is scientific method ap-

> plied self-consciously to problems more general than those ordinarily considered within a particular science. [It is] self-conscious science. (Harman, 1967, p. 343)

(D) *Analytic continuity*: Naturalizing epistemology typically involves establishing analytic continuity between the *language* of epistemology and that of science. I interpret analytic continuity broadly here to include analysing, explicating, reducing, and stating criteria for epistemological concepts, predicates, or assertions in terms of purely descriptive concepts, predicates, or assertions. For example, one creates analytic continuity between epistemology and science by proposing reliability as a conceptual analysis, reductive definition, or criterion of epistemic warrant. Naturalists oppose those like Chisholm (1977) who reject such analyses, no matter how weak. Defending some form of analytic continuity are: Armstrong (1973), Boyd (1982), Devitt (1984), Dretske (1981), Goldman (1967, 1978a, 1979a, 1981, 1986), Kim (1988), Kitcher (1983), Kornblith (1980), Lycan (1988), Schmitt (1983, 1984), Sosa (1980a, 1980b) and Swain (1981, 1985). I return to analytic continuity in more depth below.

(E) *Metaphysical continuity*: Naturalizing epistemology also typically involves establishing metaphysical continuity between the *object domains* of (i.e., objects studied by) epistemology and science. Epistemic properties, facts, or states of affairs are argued to be identical with, constituted by, or supervenient upon descriptive properties, facts, or states of affairs. Epistemic value is anchored to descriptive fact, no longer entering the world autonomously as brute, fundamental fact (as (Chisholm, 1977) suggests). I explore metaphysical continuity further below.

(F) *Axiological continuity*: Many naturalists also claim that inquiry into epistemic ends and norms must be conducted a posteriori. The suitability of justification rules is to be evaluated empirically in terms of their instrumental utility promoting epistemic ends. Epistemic ends are grounded in facts about ourselves (e.g., our contingent desires, what it is for us to desire something, etc.), our environment, and how our actions affect our ability to realize our desires. Whether high truth ratio, empirical adequacy, or pragmatic efficacy are worthwhile or desirable epistemically speaking is determined a posteriori. Epistemic ends are not given a priori by reflecting upon the nature of rationality or intellectual responsibility. Those embracing axiological continuity include: Dewey (1939a, 1958, 1960), Fuller (1988), Hooker (1987), Laudan (1984a, 1987a, 1987b, 1989) and Putnam (1981).

III. Varieties of Metaphysical and Analytic Continuity

Naturalists disagree over how to interpret the continuity uniting epistemology and science. I examine two key areas of disagreement: metaphysical-analytic and contextual-epistemological-methodological. I begin with the former.[7]

Epistemology and science are typically seen as falling on opposite sides of the fact *vs.* value divide; epistemology on the latter, science, the former. Epistemology—insofar as it is concerned with epistemic value, justification, and the regulation of belief—is viewed as a normative enterprise. It plays a critical-evaluative role in our affairs: we turn to it for the correct standards of worth from the standpoint of maximizing epistemic goals. It plays a regulative-practical role: we look to it for advice concerning how we *ought* to conduct inquiry or what we *ought* believe. Epistemic judgments, concepts, and properties are essentially practical or action-oriented. Science is seen as a theoretical enterprise concerned with rendering the world. Scientific assertions, concepts, and properties are essentially descriptive. Science tells us what *is* the case—not what *ought* to be the case.

Attempts to integrate epistemology into science run afoul of this picture. Can naturalized epistemology continue to function normatively in our affairs? Or does it become a wholly positive enterprise, stripped of all powers to regulate or criticize? Naturalists disagree. Yet continuity per se is not at issue; rather, the nature of the continuity is. Naturalists disagree over the proper strategy for bridging the fact vs. value gap and constructing metaphysical-analytic continuity between science and epistemology. They standardly pursue one of three strategies: (a) eliminativism denies epistemic value, replacing it with descriptive fact; (b) realism "firms up" epistemic value by making it depend upon facts; (c) irrealism "softens up" facts by making them depend upon epistemic value. Realists and irrealists differ over the *direction* of the dependency binding fact and value; eliminativists deny value in favor of facts.

(A) *Eliminativist strategies*: Eliminativists deny

the existence of normative epistemic properties, forsaking the standard account of knowledge as justified true belief in favor of a purely descriptive account. The normative dimension of knowledge is abandoned; stock-in-trade normative notions like justification, reasons, and evidence play no role in their epistemologies. Normative epistemology plays no useful role in our theory of the world and drops out, replaced tout court by descriptive endeavors like cognitive psychology, linguistics, or neuroscience. Epistemology becomes a wholly descriptive enterprise concerned with the causal explanation of belief. It no longer plays a critical-evaluative role *vis-a-vis* belief; it neither prescribes nor lays down rules for the proper regulation of doxastic behavior.

Quine (1969a, 1974, 1975a), of course, represents the *locus classicus* of eliminativism. Epistemology describes the causal-nomological relations linking meagre sensory input and torrential theoretical output. Stock normative epistemic notions play no obvious role in the analyses of knowledge proposed by Armstrong (1973), Goldman (1967, 1978), and Nozick (1981). Dretske (1985, p. 177) eliminates what he calls "the philosopher's usual bag of tricks (justification, reasons, evidence...)," reducing knowledge via theoretical bridge principles to "information-produced belief" (1981, p. 87). Campbell (1974) characterizes evolutionary epistemology as "descriptive epistemology, descriptive of man as a knower;" Bloor and Barnes' strong program in the sociology of knowledge defines knowledge as "any collectively accepted system of beliefs" (1982, p. 22). Other eliminativist projects include: Patricia Churchland (1986), Paul Churchland (1979, 1981), Roth (1983), and Stich (1983).

Eliminativism succeeds in integrating epistemology into science but at the cost of normativity. Epistemology appears to suffer the fate of morality and aesthetics under logical positivism: epistemic judgments acquire cognitive significance as descriptive judgments of psychology, etc., or else lose cognitive significance altogether. It seems to leave us with no resources for making cognitively significant normative judgments.

(B) *Realist strategies*: Realists seek an account of epistemic value which preserves the central evaluative functions of epistemology while permitting them to be carried out within the context of science. Epistemic judgments are cognitively significant and capable of bearing truth values in a fundamentally non-normative sense. Epistemic predicates are genuinely referential. Epistemic properties (like warrant) are objective and defined in non-normative terms such as verific reliability (truth-conduciveness), pragmatic reliability (successful prediction and control), or reproductive fitness. There is a theory- and evidential-practice-independent fact of the matter about the epistemic status of our cognitive activity. A variety of realist views are defended.

(i) *Criterialism*: Criterialists claim epistemic properties supervene upon but are neither identical with nor constituted by descriptive properties. Epistemic properties are metaphysically determined by naturalistic ones without epistemic predicates or concepts being defined or reduced to naturalistic ones. Epistemic properties E supervene upon natural properties N just in case: necessarily, for any belief p and E-property J, if p has J then there is a N-property R such that (i) p has R and (ii) necessarily, whatever has R has J.[8] Every instantiated epistemic property must have a sufficient (i.e., entailing) condition among its natural properties. Epistemic properties (like justification) obtain in virtue of or because of non-epistemic ones (like verific reliability). They are "fixed" by natural base properties and are consequently no longer autonomous from the natural. Epistemic properties remain essentially normative. Although reduction is generally viewed as a special case of supervenience, criterialists deny the reducibility of the epistemic. Epistemic predicates cannot be naturalistically defined or reduced without loss of their essential normative content. We can thus specify only naturalistic *criteria* or conditions for them. Criterialists include Goldman (1979a, 1981, 1986), Kim (1984, 1988) and Sosa (1980a, 1980b).

With the help of naturalistic criteria such as provided Goldman's historical reliabilism, normative epistemology makes use of factual, aposteriori considerations in evaluating belief. It need appeal to no special cognitive faculties. Criterialism thus succeeds in preserving the central normative role of epistemology while avoiding what many see as the excesses of eliminativism, reductionism, or definism, on the one hand, and non-naturalism (e.g., Chisholm, 1977), on the other. The former naturalize epistemology at the cost of its normativity; the latter preserves normativity but divorces epistemology from the natural. For criterialists epistemic properties remain sui generis since essentially normative—their determination by naturalistic base properties

notwithstanding. They are fixed by descriptive properties via supervenience relations yet remain distinctly normative. In short, supervenience preserves normativity while denying autonomy.

(ii) *Definism*: Definists construe epistemology as essentially theoretical and descriptive. Epistemic terms, notions, or judgments are defined (intensionally) without remainder in descriptive terms. Epistemic properties, facts, or states of affairs are identified with descriptive ones. For example, Maffie (1988, f) defines justification in terms of undefeated reliability (which is defined realistically in terms of truth). Epistemic judgments assess the fitness of cognitive behavior relative to epistemic ends i.e. the likelihood of a belief's being true given its causal ancestry. Epistemic judgments resemble ordinary instrumental judgments of means-ends relations such as an engineer's evaluating the suitability of a steel alloy for constructing a railway bridge. They attribute descriptive properties to cognitive behavior and are verifiable scientifically. Epistemic value (e.g., warrant) is a descriptive fact about a particular species of means-ends relationship in the world. Epistemic "oughts" are identical with descriptive facts about instrumentally appropriate behavior relative to epistemic ends. Instead of eliminating stock epistemic properties like justification, definism identifies them with descriptive ones. Finally, such definitions are offered as a *reforming* definitions of our ordinary, folk epistemic expressions. As instances of "semantic ascent" in the course of ordinary scientific theory construction, they are warranted to the degree they contribute to our overall scientific picture of the world and human cognizers (including our epistemic intuitions, judgments, or practices).[9]

(iii) *Reductionism*: Nagel-style (1961) reduction enables naturalists to integrate epistemology within science by reducing epistemology to science. It establishes continuity but unlike eliminativism, it legitimizes rather than eliminates epistemic properties. Epistemology reduces to science just in case the primitive concepts or predicates of epistemology are extensionally (rather than intensionally as with definism) definable in terms of those of natural science. The terms of normative epistemology are connected with those of science by "bridge laws" or nomological correspondence rules, the result being that the properties expressed by the predicates of normative epistemology are nomologically co-extensive with those expressed by the non-normative predicates of natural science. These reductive definitions are synthetic and aposteriori. Epistemic properties no more cease to exist upon reduction, however, than do tables upon reduction to bundles of atoms or water upon reduction to water. Epistemic properties may undergo two sorts of reduction: they may be reductively identified with (e.g., as in "water = H_2O") or reductively constituted by (e.g., as in "table = bundle of atoms") descriptive properties.

According to Lycan (1988, p. 144), epistemic notions reduce to the "teleological notions of the theory of organ systems" which are standardly employed by biologists and biochemists. Epistemic goodness, justification, and so on are defined in terms of biological utility i.e., reproductive fitness. Epistemic value is a species of descriptive, instrumental value. The epistemic goodness of a cognitive process is a matter of its goodness-of-design relative to reproductive fitness. Epistemic value judgments are a species of descriptive judgments concerning the efficient promotion of biological utility. The normative force of epistemic terms and judgments derives from "value notions implicit or explicit in design-stance psychology" (1988, p. 143). Our epistemic methods are *desirable* because they promote reproductive fitness.

Devitt (1984), Swain (1981, 1985) and Schmitt (1984) also propose reductive views. Unlike Lycan, however, they conceive epistemic justification in terms of truth.[10]

Can definism and reductionism preserve the important normative role of epistemology in our lives? Laudan (1987a, 1987b, 1989) and Maffie (1988, f) supply the requisite machinery. Epistemology is essentially descriptive and only hypothetically normative. It enjoys an intimate relationship with human motivation and conduct—and is therefore normative—in virtue of its centrality and widespread utility as a means to our contingent, variable ends. It is by and large rational for us to heed the epistemic. Yet this intimacy and hence normativity are contingent and hypothetical only. Like other descriptive endeavors, epistemology becomes normative only within the framework of instrumental reason and its normativity is thus parasitic upon that of the latter. Epistemic concepts, judgments, or facts become reason-giving, action-oriented, or attitude-molding only when relevant to our ends. Epistemic rules are

conditional—rather than categorical—imperatives which are empirically defeasible, "contingent claims about optimal ways to realize our ends" (Laudan, 1987b, p. 225). They report on the contingent suitability of means to ends and acquire normative force only when situated within a context of appropriate human end-seeking. In sum, the normativity of epistemology is grounded in instrumental reason together with contingent facts about ourselves (our contingent ends and motivational make-up), our environment, and what we must do to realize our ends in that environment. Epistemology plays a central normative role in our lives in virtue of its instrumental utility relative to the satisfaction of our variable, contingent ends.[11]

(C) *Irrealist strategies*: Naturalist irrealists (e.g., Dewey (1960), Ellis (1985, 1988) and Putnam (1981)) reverse the "direction" of metaphysical-analytic continuity linking epistemology and science.[12] Unlike realists, they define the non-normative in terms of the normative. Descriptive properties, facts, or states of affairs are identified with or reduced to normative ones. "A fact is something that it is rational to believe, or, more precisely, the notion of a fact (or true statement) is an idealization of the notion of a statement that is rational to believe," writes Putnam (1981, p. 201). In short, the facts are "soft all the way down." Epistemic properties are defined in terms of normative properties such as moral goodness or rationality. A cognizer's being justified depends of human evidential judgments or practices. Ellis (1988, pp. 431-433) argues that the way the world is depends upon our epistemic perspective which is defined by our system of values. Truth is what it is right to believe (1988, p. 426). For Putnam, truth is ideal rational assertability. Descriptive notions, judgments, or predicates are also at bottom normative. In sum, facts are derivative, values basic; the relevant values are *our* values, grounded in our environment and our make-up qua human beings.

In closing, what is the source of epistemolgy's normativity? Eliminativists posit no source since they deny epistemology's normativity. Irrealists as well as reductionist and definist realists base normativity upon contingent facts about our ends, motivational make-up, environment, and how that environment affects our reaching our ends. Criterialists appeal to essentially normative, sui generis states of affairs or properties.[13]

IV. The Scope of Continuity

Naturalists disagree over how *far* naturalizing epistemology should go. At issue is the *scope* of contextual, methodological, and epistemological continuity.

Like ethics, epistemology admits of meta- and normative levels. Normative epistemology determines the scope and sources of human knowledge, evaluates and prescribes doxastic behavior, etc. Meta-epistemology studies the semantics and validation of epistemic claims. It determines the cognitive and epistemic status of epistemic judgments, the reference of epistemic expressions, etc. Meta-epistemology also includes axiological inquiry into the nature and content of epistemic ends and norms. Naturalists uniformly advocate naturalizing normative epistemology; but there consensus disappears.

According to what I call "unlimited naturalism," epistemic inquiry is to be naturalized "all the way up" i.e., including meta-epistemology itself. Determining the proper ends of inquiry from the epistemic point of view, the correct theory or rules of epistemic justification, or the cognitive status of epistemic assertions, etc. are aposteriori activities to be addressed from within science as a part of constructing our best overall theory of the world and our place within it. There are no autonomous, meta-epistemological problems requiring special non-empirical methods. Claims about the aims and nature of knowledge are no different in kind from ordinary scientific claims. Epistemic ends are not given apriori by reflection upon the nature of rationality or intellectual responsibility but rather by aposteriori reflection upon descriptive facts about ourselves (e.g., our contingent desires, what it is for us to desire something, etc.), our environment, and how our actions affect our ability to realize our desires. The suitability of high truth ratio, reproductive fitness, etc. as epistemic ends is assessed in terms of facts about the world together with our other ends.

> Which epistemic goals it is rational to choose [is] a function of our nature and circumstances on the one side and critical reflection on the process of acquiring knowledge (in light of historical experience and theory) on the other. (Hooker, 1987, 12f.)

Justification rules are defeasible and corrigible conjectures about means-ends relationships and are evaluated empirically in terms of their instrumental

utility promoting epistemic ends. Whether or not they promote epistemic ends is a contingent matter which depends on the contingent structure of the world. Epistemic theories are scientific hypotheses about ways of acting in the world and are subject to the same techniques of adjudication as brought to bear upon ordinary scientific theories. In short, inquiry into ends, norms, and facts are "cut from the same cloth" (Laudan, 1989).

Epistemology and science are both experimental enterprises displaying what Boyd (1980, 1984) calls a "dialectical relationship of mutual dependence." They are mutually accomodating, mutually adjusting, each being subject to critical feedback from the other. What appears to be a circle, argues Hooker (1974, p. 415), is really "a spiral extending along the historical time axis: epistemological theory evaluates scientific method, scientific discovery informs epistemological theory." Spinoza likened the process to the technological development of the hammer. On the basis of the performance of an earlier hammer we construct using that selfsame hammer, a subsequent, more effective hammer, and so on. Feedback operates upon both norms and substantive claims. For example, Laudan (1977, 60ff.) argues that the classic norms of inductivist methodology were later revised in light of the instrumental success of so-called "theoretical" entities which were considered scientifically illegitimate by these norms. Confronted with a conflict between normative principles on the one hand, and the instrumental utility of theoretical entities on the other, scientists chose to revise the classic principles of induction. They were eventually replaced by the principles of the hypothetico-deductive method. In short, we assess norms in light of practical results, and assess practical results in light of norms. Unlimited naturalists have much to learn from Rescher's (1977) "systems-theoretic" approach to epistemology. While not obviously naturalistic, Rescher's approach nevertheless provides a detailed and ingenious account of the interdependence of norms, methods, and substantive claims sought by unlimited naturalists.[14]

Unlimited naturalists reject stock philosophical methods e.g., conceptual analysis, reflective equilibrium, or intuitionism as non-naturalistic and of dubious epistemic merit. Stich (1988) claims there is compelling empirical evidence showing these to be non-truth-conducive and unsuited to confer epistemic warrant. Fuller (1989) attacks both the contingent desirability and feasibility of reflective equilibrium. Boyd (1984) doubts there are any significant analytic or conceptual truths about any scientifically interesting subject matter. In sum, I suggest we view unlimited naturalism as trying to "fit epistemology into science" and create a thorough-going, scientific epistemology.

According to "limited naturalism," epistemic, methodological, and contextual continuity are confined to normative level epistemological inquiry (e.g., evaluating or prescribing belief). However, meta-epistemological level inquiry into the correct theory or norms of justification, the meaning of epistemic expressions, and axiological inquiry into the proper goals of cognition from the epistemic point of view are conducted by extra-scientific, distinctly philosophical methods. Scientific methods are unsuited for such matters and consequently significant portions of epistemic inquiry remain autonomous from scientific investigation. Aposteriori considerations are irrelevant to "foundational" questions in epistemology, claims Goldman (1986, p. 9). Facts about cognitive faculties are relevant to evaluating doxastic behavior but not to choosing the correct criterion of *J*-rules or to analysing our ordinary notion of justification. The latter involve reflective equilibrium.

> This procedure need not involve empirical psychology or social science. I do not claim that psychology plays a role in selecting a criterion of *J*-rule rightness. This is not the level at which psychology enters the epistemological enterprise. It enters the picture only if and when a criterion is selected that makes reference to cognitive processes.[15]

Goldman (1978a, 1979, 1986), Sosa (1980a, 1980b), Kim (1988) and Swain (1981, 1985) defend theories of epistemic ends, value, or norms using non-empirical methods such as conceptual analysis, explication of ordinary usage, appeal to intuitions, or reflective equilibrium. In sum, I suggest we view limited naturalism as merely trying to "fit science into epistemology."

V. Criterialism Reconsidered

Criterialism appears to offer realists an extremely attractive strategy for naturalizing epistemology. Supervenience preserves the normativity while denying the metaphysical autonomy of epistemology, enabling criterialists to capture the virtues of both

naturalism and non-naturalism. Unfortunately, however, criterialism preserves normativity by leaving the fact/value distinction intact, denying the justificational and methodological continuity of science and epistemology, and removing epistemic value from our daily experience.

Let's begin by asking how criterialism differs from non-naturalist views and whether or not it actually succeeds in naturalizing epistemology. It does not succeed by virtue of advocating supervenience alone seeing as paradigmatic non-naturalists like Chisholm (1982, pp. 50-60) in epistemology and Moore in ethics embrace the supervenience of value upon non-value. Nor does it succeed in virtue of specifying descriptive criteria for evaluative notions; Chisholm and Moore are willing to do this, too. Criterialism views epistemic properties, notions and judgments as sui generis or essentially non-descriptive. Why, then, characterize criterialism as a naturalism at all?

If we view naturalism as first and foremost an *ontological* monism admitting of only one kind of fact—*viz.* descriptive, then criterialism does not qualify as a naturalism. For example, were we to equate naturalism with physicalism and conceive physical properties as essentially descriptive, then criterialism's sui generis, normative properties would not be admissible as natural properties. However, seeing as most naturalists eschew physicalism, these grounds seem inadequate for disqualifying criterialism.

If we follow Dewey (1939b), Danto (1967), Randall (1944) and Randall and Buchler (1942) and view naturalism as first and foremost a *methodological* monism admitting of only one kind of cognitive method—*viz.* scientific, then criterialism *appears* to qualify as a kind of naturalism. It provides access to epistemic properties via scientific methods—the sui generis, normative character of these properties notwithstanding. Armstrong (1978) views naturalism as tolerating ontological diversity so long as what exists is integrated within the natural order, cognitively accessible via scientific methods, and causally efficacious. Naturalism excludes only what is beyond the pale of scientific methodology. Therefore, if supervenience figures normative epistemic properties into the natural order, making them neither *epistemically* sui generis (i.e., requiring special, non-scientific methods or evidence) nor *metaphysically* sui generis (i.e., autonomous), then they are naturalistically respectable and criterialism qualifies as a species of naturalism.

But criterialism fails to meet Armstrong's conditions. It conceives epistemic properties as possessing a sui generis normative residue which in principle eludes apprehension via scientific methods and which is therefore in principle beyond the methodological-epistemological pale of naturalism. In short, criterialists are limited naturalists. Meta-level inquiry into the nature of epistemic ends, norms, and value, the meaning and validation of epistemic judgments, etc. is epistemically and methodologically *discontinuous* with science, requiring instead special, apriori, non-natural methods. Here lies the rub.

Thus our dissatisfaction with criterialism need not be based on a narrow-minded commitment to physicalism but rather only on a thorough-going commitment to the continuity of epistemology and natural science. Criterialism affirms the epistemological sui generis character of the epistemic, leaving us with an epistemological dualism of facts and values. Epistemology is still largely prior to and autonomous from science since it still employs non-natural cognitive methods. The consequent bifurcation of human inquiry undermines the integrity of naturalism as a comprehensive methodological and epistemological program.

Criterialism faces additional troubles raised by the following dilemma. On the one hand, if sui generis, essentially normative epistemic properties are causally efficacious, then they are, as Mackie (1977,38) points out, ontologically "queer" i.e., "of a very strange sort, utterly different from anything else in the universe." For unlike other natural properties, they have "to-be-pursuedness" (1977, p. 40) built into them since they are categorically prescriptive and intrinsically motivating. On the other hand, if they are causally inefficacious then they are removed from human cognition. Without causal efficacy and causal mechanisms for learning epistemic facts, how do we gain epistemic access to them? In attempting to preserve epistemolgy's normativity, criterialism (like non-naturalism) removes epistemic value from our daily lives. What's more, without causal effect on the world, these properties seem otiose and Occam's razor enjoins us not to posit them. In either case, criterialism exacts serious costs.

But perhaps these costs are worth paying since criterialism represents realists' only hope for preserving the normativity of epistemology? It doesn't.

As we saw earlier, reductionists and definists may preserve normativity without appealing to sui generis, non-descriptive states of affairs or the special cognitive methods needed for their apprehension.

VI. Conclusion

Continuity lies at the heart of the naturalistic turn in epistemology. Yet the heart is not unvexed. Limited naturalism with its autonomous meta-epistemology and axiology attracts those bent on maintaining the essential normativity of epistemic ends, language, or properties. Scientific methods seem ill-equipped for these, so special non-natural methods are introduced. The lesson to be learned from criterialism is that epistemic inquiry into natural properties demands that they be not only metaphysically continuous but also causally accessible. Otherwise, there will be a price to pay down the road in terms of epistemological discontinuity. This causal condition on cognitive access does not represent an apriori pronouncement of the sort issued by verificationists; rather, the hardwon fruit of empirical research into human cognition. Epistemological, metaphysical and axiological continuity must be in harmony if we are to have a coherent naturalistic picture of the world as well as human end-seeking and activity within it.

North Carolina State University

Received September 11, 1989

NOTES

1. I'd like to thank Steve Fuller, Allan Gibbard, Julia Greene, Larry Laudan, Louis Loeb, Bonnie Paller, Peter Railton, Paul Roth, Lawrence Sklar, David Stump, and Stephen Yablo for their helpful comments and criticisms at various stages of this essay. I also benefitted from discussions in Larry Laudan's 1989 NEH Summer Seminar on "Naturalized Epistemology" during which it was completed. At the end of the essay is a bibliography of works cited.

2. Naturalists disagree over the culpability of elements in addition to (1) and (3). For example, naturalist irrealists (e.g., Dewey (1960), Ellis (1988), and Putnam (1981)) reject the realist notions of truth and existence expressed in elements (2) and (4); while naturalist realist s (e.g., Boyd (1980, 1984, 1985), Hooker (1987), and Devitt (1984)) uphold these.

3. On Piaget, see Kitchener (1980, p. 133). Hooker (1987, p. 261) urges us to naturalize epistemology on the grounds that the history of science shows that naturalism is "the most theoretically fecund methodology available."

4. Naturalism opposes a variety of conceptions of epistemology which deny continuity in one or more of these six ways. Time does not allow my exploring these here. In what follows I use "epistemic" and "epistemological" interchangeably. Helpful discussions of naturalism include:Alston (1978), Armstrong (1978), Danto (1967), Edel (1946), Kim (1988), Krikorian (1944), Murphy (1945), Randall (1944), Randall and Buchler (1942), and Sheldon (1945).

5. Goldman (1978b, p. 522). See also: Armstrong (1978), Boyd (1984), Devitt (1984), Friedman (1979), Goldman (1980), Giere (1985), Hooker (1987), Laudan (1987a, 1987b, 1989), and Rescher (1977).

6. Piaget's criticisms of the factual presuppositions of Humean, Kantian, and logical positivist epistemologies are discussed in Kitchener (1980a, 1980b, 1986). Hooker (1974, p. 413) writes, "recent advances in physics have destroyed the synthetic *a priori* of Kantian space" and "recent research in perception has rendered ludicrous the notion of pre-occurring conceptualized data." See also: Fuller (1989), Goldman (1986), Kornblith (1988), and Stich (1988).

7. Two areas of dispute which I cannot explore here are: (a) the scientific methods epistemology may legitimately employ: e.g., Boyd, Devitt, and Hooker include abduction, Fine and Van Fraassen do not; and (b) the preferred scientific discipline(s) with which to wed epistemology: e.g., evolutionary biology (Campbell), neuroscience (Patricia Churchland, Paul Churchland), behaviorist psychology and linguistics (Quine), cognitive sociology (Fuller, Bloor, and Barnes), genetic psychology (Piaget), cognitive psychology (Goldman), history of science (Laudan); or ethnomethodology (Latour and Woolgar, 1979)

8. Kim (1978, 1984, 1988) distinguishes weak from strong forms of supervenience and defends strong supervenience as appropriate for value theory. Since criterialists share Kim's conclusion, I limit discussion to strong supervenience.

9. For further discussion of semantic ascent, see Quine (1960, pp. 258f.).

10. The sort of reduction—identity or constitution—proposed by Lycan, Devitt, Schmitt, and Swain is not clear.

11. Judgments of instrumental rationality are essentially normative since defined in terms of adopted ends. Because they describe means to our ends and the best view of what it is to have an end is to be motivated to take steps towards reaching it, judgments of rationality provide us with irresistible reasons for acting. Motivation is built into the rational because what is rational is defined in terms of *our* ends.

12. Interpreting Putnam as a naturalist may seem controversial in light of his (1983a) claim that reason cannot be naturalized. However, Putnam's target there is "monistic naturalism" or physicalism, as he makes clear in (1981, p. 211). Putnam rejects realist or descriptivist varieties of naturalism—not naturalism per se.

13. There seems to be an affinity between one's choice of strategy and one's reason for championing science as the appropriate context for conducting post-Cartesian epistemology. Eliminativists and irrealists appear to have at their disposal pragmatic reasons only. Science is a refinement of ordinary, common sense methods of inquiry, being a more effective procedure for realizing ends (whatever they be). Its claim to our doxastic loyalty rests upon its greater problem-solving ability and our wanting our problems solved. In contrast, realists may also be scientific realists, championing science on grounds of its superior epistemic status i.e., its greater truth-conduciveness. Science is better than its rivals because it yields a true (or at least more approximately true) picture of the world. Its claim to our doxastic loyalty rests upon its quite specific epistemic success at correctly capturing external reality.

14. See also Laudan's (1984) "reticulational model of scientific rationality" and Quine's (1969, p. 83) "reciprocal containment" of epistemology and science. Other unlimited naturalists include Dewey, Fuller, Giere, Maffie, and Putnam.

15. Goldman (1986, p. 66). On p. 181, Goldman writes, "The principal way cognitive science can contribute to epistemology...is to identify basic belief-forming, or problem-solving, processes. Once identified, these processes would be examined by primary epistemology according to the evaluative dimensions and standards justified by reflective equilibrium. See also Goldman (1978a, p. 520), where he speaks of "infusing psychology into epistemology."

BIBLIOGRAPHY

Alston, William. (1978), "Meta-Ethics and Meta-Epistemology," in A. Goldman & J. Kim (eds.), (1978), pp. 275-97.

Armstrong, D.M. (1973), *Belief, Truth, and Knowledge* (Cambridge: Cambridge University Press).

Armstrong, D.M. (1978), "Naturalism, Materialism, and First Philosophy," *Philosophia,* vol. 8, pp. 261-76.

Asquith, P.D., and Giere, R.N., (eds.). (1982) *PSA* 1980, vol. 2 (East Lansing: Philosophy of Science Association).

Barnes, Barry, and Bloor, David. (1982), "Relativism, Rationality and the Sociology of Knowledge," in Hollis and Lukes (eds.), (1982), pp.21-47.

Bloor, David. (1976), *Knowledge and Social Imagery* (London: Routledge & Kegan Paul).

Boyd, Richard. (1982), "Scientific Realism and Naturalized Epistemology," in Asquith and Giere (eds.), (1980), pp. 613-62.

Boyd, Richard. (1984), "The Current Status of Scientific Realism," in Leplin (ed.), (1984), pp. 41-82.

Campbell, Donald T. (1974), "Evolutionary Epistemology," in Schilpp (ed.) (1974), pp. 413-63.

Chisholm, Roderick. (1977), *The Theory of Knowledge* 2nd ed. (Englewood Cliffs: Prentice-Hall).

Chisholm, Roderick. (1982), *The Foundations of Knowing* (Minneapolis: University of Minnesota Press).

Copp, D., and Zimmerman, D.(eds.). (1984), *Morality, Reason, and Truth* (Totowa,N.J.: Rowman & Littlefield).

Churchland, Patricia. (1986), *Neurophilosophy: Toward a Unified Science of the Mind/Brain* (Cambridge, Mass: MIT Press).

Churchland, Paul. (1979), *Scientific Realism and the Plasticity of Mind* (Cambridge: Cambridge University Press).

Churchland, Paul. (1981), "Eliminative Materialism and the Propositional Attitudes," *The Journal of Philosophy*, vol. 78, pp. 67-90.

Churchland, Paul, and Hooker, Clifford A. (eds.). *Images of Science* (Chicago: University of Chicago Press).

Danto, Arthur C. (1967), "Naturalism," in Edwards (ed.), (1967), vol. 5, pp. 448-50.

Davidson, Donald, and Hintikka, Jaakko (eds.). (1975), *Words and Objections: Essays on the Work of W.V. Quine* (Dordrecht, Holland: Reidel).

Devitt, Michael. (1984), *Realism and Truth* (Princeton: Princeton University Press).

Dewey, John. (1939a), *Theory of Valuation.* International Encyclopedia of Unified Science, vol.2, no.4, O. Neurath, R. Carnap, & C. Morris (eds.) (1939).

Dewey, John. (1939b), "Experience, Knowledge, and Valuation: A Rejoinder," in Schilpp (ed.) (1939), pp. 515-608.

Dewey, John. (1948), *Reconstruction in Philosophy* (Boston: Beacon Press).

Dewey, John. (1958), *Experience and Nature* (New York: Dover Publications).
Dewey, John. (1960), *The Quest For Certainty* (New York: G.P. Putnam's Sons).
Dretske, Fred. (1981), *Knowledge and the Flow of Information*, (Cambridge: MIT Press).
Edel, Abraham, (1946), "Is Naturalism Arbitrary?" *Journal of Philosophy,* vol. 43, pp.141-52.
Edwards, Paul. (1967), *The Encyclopedia of Philosophy* (New York: Macmiillan & Free Press).
Ellis, Brian. (1985), "What Science Aims to Do," in P. Churchland and C.A. Hooker (eds.), (1985), pp.75-98.
Ellis, Brian. (1988), "Internal Realism," *Synthese,* vol. 76, pp. 409-34.
Fine, Arthur. "The Natural Ontological Attitude," in Leplin (ed.), (1984), pp.83-107.
French, Peter A. Uehling, Theodore E., Jr. and Wettstein, Howard K. (eds.), (1980), *Studies in Epistemology*. Midwest Studies in Philosophy, vol. 5, (Minneapolis: University of Minnesota Press).
French, Peter A. Uehling, Jr. Theodore E. and Wettstein, Howard K. (eds.), (1981), *Analytic Philosophy*. Midwest Studies in Philosophy, vol.6., (Minneapolis: University of Minnesota Press).
Friedman, Michael. (1979), "Truth and Confirmation," *Journal of Philosophy,* vol. 76, pp. 361-82.
Fuller, Steve. (1988), *Social Epistemology* (Bloomington: Indiana University Press).
Fuller, Steve. (1989), *Philosophy of Science and Its Discontents* (Boulder: Westview Press).
Giere, Ronald N. (1985), "Philosophy of Science Naturalized," *Philosophy of Science,* vol. 52, pp. 331-56.
Goldman, Alvin. (1967), "A Causal Theory of Knowing," *Journal of Philosophy,* vol. 64, pp. 357-72.
Goldman, Alvin. (1978a), "Discrimination and Perceptual Knowledge," in Pappas and Swain (eds.), (1978), pp. 120-54.
Goldman, Alvin. (1978b), "Epistemics: The Regulative Theory of Cognition," *Journal of Philosophy*, vol. 75, pp. 509-23.
Goldman, Alvin. (1979), "What Is Justified Belief?" in Pappas (ed.), (1979), pp. 1-23.
Goldman, Alvin. (1980), "The Internalist Conception of Justification," in French, et.al. (eds.), (1980), pp. 27-52.
Goldman, Alvin, (1981), "Review: *Philosophy and The Mirror of Nature* by Richard Rorty," *Philosophical Review,* vol. 90, pp.424-29.
Goldman, Alvin. (1986), *Epistemology and Cognition* (Cambridge: Harvard University Press).
Goldman, A.I., and Kim, J., (eds.). (1978), *Values and Morals* (Dordrecht: Reidel).
Guttenplan, Samuel (ed.). (1975), *Mind and Language* (Oxford: Clarendon Press).
Haack, Susan. (1975), "The Relevance of Psychology to Epistemology,"*Metaphilosophy,* vol. 6, pp.161-176.
Harman, Gilbert. (1967), "Quine on Meaning and Existence: II," *Review of Metaphysics,* vol. 21, pp.343-367.
Hollis, Martin, and Lukes, Stephen (eds.). (1982), *Rationality and Relativism* (Cambridge. Mass.: MIT Press).
Hooker, C.A. (1974), "Systematic Realism," *Synthese*, vol. 26, pp.409-97.
Hooker, C.A. (1987), *A Realist Theory of Science* (Albany: State University of New York Press).
Kim, Jaegwon. (1978), "Supervenience and Nomological Incommensurables," *American Philosophical Quarterly* 15, pp. 149-56.
Kim, Jaegwon. (1984), "Concepts of Supervenience," *Philosophy and Phenomenological Research,* vol. 65, pp. 153-76.
Kim, Jaegwon. (1988), "What Is 'Naturalized Epistemology?'," in Tomberlin (ed.), (1988), pp. 381-406.
Kitchener, Richard F. (1980a), "Piaget's Genetic Epistemology," *International Philosophical Quarterly,* vol. 20, pp. 377-405.
Kitchener, Richard F. (1980b), "Genetic Epistemology, Normative Epistemology, and Psychologism, *Synthese*, vol. 45, pp. 257-80.
Kitchener, Richard F. (1986), *Piaget's Theory of Knowledge: Genetic Epistemology and Scientific Reason* (New Haven: Yale University Press).
Kitcher, Phillip. (1983), *The Nature of Mathematical Knowledge* (New York: Oxford University Press).
Kornblith, Hilary. (1988), "How Internal Can You Get?" *Synthese,* vol. 74, pp. 313-27.
Kornblith, Hilary (ed.). (1985), *Naturalizing Epistemology* (Cambridge MIT Press).
Krikorian, Yervant H. (ed.). (1944), *Naturalism and the Human Spirit* (Morningside, NY: Columbia University Press).
Laudan, Larry. (1977), *Progress and Its Problems* (Berkeley: University of California Press).
Laudan, Larry. (1984a), *Science and Values* (Berkeley: University of California Press).
Laudan, Larry, (1984b), "A Confutation of Convergent Realism," in Leplin (ed.) (1984), pp. 218-49.
Laudan, Larry. (1987a), "Progress or Rationality? The Prospects for Normative Naturalism," *American Philosophical Quaterly*, vol. 24, pp.19-31.
Laudan, Larry. (1987b), "Relativism, Naturalism and Reticulation," *Synthese*, vol. 71, pp. 221-34.
Laudan, Larry. (1989), "Normative Naturalism," *Philosophy of Science* (forthcoming).
Latour, Bruno, and Woolgar, Steve. (1979), *Laboratory Life* (London: Sage).

Leplin, Jarrett. (1984), *Scientific Realsim* (Berkeley: University of California Press).
Lycan, William. (1988), *Judgment and Justification* (Cambridge: Cambridge University Press).
Mackie, J. L. (1977), *Ethics: Inventing Right and Wrong* (New York: Penguin Books.
Maffie, James. (1988), *The Nature and Province of Naturalized Epistemology* Ph.D Dissertation, University of Michigan.
Maffie, James, (forthcoming), "Naturalism and the Normativity of Epistemology," *Philosophical Studies.*
Murphy, Arthur E. (1945), "Review of *Naturalism and the Human Spirit*," *Journal of Philosophy,* vol. 42, pp. 400-17.
Nagel, Ernest. (1961), *The Structure of Science* (New York: Harcourt, Brace, & World).
Neurath, Otto; Carnap, Rudolph; Morris, Charles, (eds.). (1939), *International Encyclopedia of Unified Science* (Chicago: University of Chicago Press).
Nozick, Robert. (1981), *Philosophical Explanations* (Cambridge: Harvard University Press).
Pappas, George S. (ed.). (1979), *Justification and Knowledge* (Dordrecht: Reidel).
Pappas, George S. and Swain, Marshall (eds.). (1978), *Essays on Knowledge and Justification* (Ithaca: Cornell University Press).
Piaget, Jean. (1971), *Insights and Illusions of Philosophy*, trans. by Wolfe Mays (New York: World Publishing Co.)
Piaget, Jean. (1972), *Psychology and Epistemology*, trans. by Arnold Rosin (New York: Viking Press).
Putnam, Hilary. (1981), *Reason, Truth, and History* (Cambridge: Cambridge University Press).
Putnam, Hilary. (1983a), "Why Reason Can't Be Naturalized," in Putnam (1983), pp. 229-47.
Putnam, Hilary. (1983b), *Philosophical Papers*, vol. 3 (Cambridge: Cambridge University Press).
Quine, W.V. (1953). *From a Logical Point of View* (New York: Harper & Row).
Quine, W.V. (1953a), "Two Dogmas of Empiricism," in Quine (1953), pp. 20-46.
Quine, W.O. (1960), *Word and Object* (Cambridge, Mass.: MIT Press).
Quine, W.V. (1969), *Ontological Relativity and Other Essays* (New York: Columbia University Press).
Quine, W.V. (1969a), "Epistemology Naturalized," in Quine (1969), pp. 69-90.
Quine, W.V. (1969b), "Natural Kinds," in Quine (1969), pp. 114-38.
Quine, W.V. (1974), *Roots of Reference* (La Salle: Open Court).
Quine, W.V. (1975a), "The Nature of Natural Knowledge," in S. Guttenplan, (ed.), (1975), pp. 67-81.
Quine, W.V. (1975b),"Reply to Smart," in Davidson and Hintikka (eds.), (1975), pp. 292-94.
Quine, W.V. (1981), "Reply To Stroud," in French, et.al. (eds.), (1981), pp. 473-75.
Randall, John Herbert. (1944), "Epilogue: The Nature of Naturalism," in Krikorian (ed.), (1944), pp. 354-82.
Randall, John Herbert, and Buchler, Justus. (1942), *Philosophy: An Introduction* (New York: Barnes & Noble).
Rescher, Nicholas. (1977), *Methodological Pragmatism* (Oxford: Basil Blackwell).
Roth, Paul. (1983), "Seigel on Naturalized Epistemology and Natural Science" *Philosophy of Science,* vol. 50, pp. 482-93.
Schilpp, Paul, (ed.). (1939), *The Philosophy of John Dewey* (LaSalle: Open Court Press).
Schilpp, Paul, (ed.). (1974), *The Philosophy of Karl Popper* (LaSalle: Open Court Press).
Schmitt, Frederick F. (1983), "Knowledge, Justification and Reliability," *Synthese*, vol. 55, pp. 209-29.
Schmitt, Frederick F. (1984), "Reliability, Objectivity and the Background of Justification," *The Australasian Journal of Philosophy,* vol. 62, pp. 1-15.
Sheldon, W.H. (1945), "Critique of Naturalism," *Journal of Philosophy,* vol. 42, pp. 253-70.
Sosa, Ernest. (1980a), "The Raft and the Pyramid: Coherence versus Foundations in the Theory of Knowledge," in French *et. al.* (eds.), (1980), pp.3-26
Sosa, Ernest. (1980b), "The Foundations of Foundationalism," *Nous*, vol. 14, pp. 547-654.
Stich, Stephen. (1983), *From Folk Psychology to Cognitive Science* (Cambridge: MIT Press).
Stich, Stephen. (1988), "Reflective Equilibrium, Analytic Epistemology and the Problem of Cognitive Diversity," *Synthese* vol. 74, pp. 391-413.
Sturgeon, Nicholas. (1984), "Moral Explanations," in D. Copp & D. Zimmerman (eds.), (1984), pp. 50-78.
Suchting, Wal. (1983), "Knowledge and Practice: Towards a Marxist Critique of Traditional Epistemology," *Science and Society,* vol. 43, pp. 2-36.
Swain, Marshall. (1981), *Reasons and Knowledge* (Ithaca: Cornell University Press).
Swain, Marshall. (1985), "Justification, Reasons, and Reliability," *Synthese*, vol. 64, pp. 69-92.
Tomberlin, James E., ed. (1988), *Philosophical Perspectives, 2: Epistemology 1988* (Atascadero, CA.: Ridgeview Publishing).
Van Fraassen, Bas C. (1980), *The Scientific Image* (Oxford: Oxford University Press).

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 27, Number 4, October 1990

CLASSIFICATION BY COMPARISON WITH PARADIGMS

Rolf Eberle

It has long been claimed that classifications with respect to "natural" properties can be both learned and justified by comparison with exemplars. Yet, work initiated especially by R. Carnap[1] and N. Goodman[2] has revealed that the construction of quality classes on the basis of various resemblance relations is beset by serious difficulties. Nevertheless, it will be shown here (by an extension of an earlier discussion[3]) that principled classifications with respect to all but universal or empty properties (and relations) can always be effected by simultaneous comparison and contrast with suitable exemplars. After an introductory section I, this will be detailed for predicates in section II (with allowance for "enumerative classifications" in section III) and for relation expressions in section IV. In section V, the advantages will be discussed of thinking of such exemplars as *paradigms*: sentences, such as "birds are feathered", in the sense of "*typical* birds are feathered", will be true of them; *vague* predicates will apply to qualitative neighborhoods around them; and *comparatives* can be justified on the basis of typical (rather than extreme) exemplars.

I. CLASSIFICATION BY COMPARISON: MOTIVATION

ACCORDING to one nominalistic tradition, abstract universals, such as as the property "white," and perhaps also classes, such as the class of all white things, do not exist. Still, classifications, such as that of all white things, are effected and justified by virtue of certain resemblance relations which exactly the white things bear to each other, plus a relation, say that of denotation, that concrete utterances or inscriptions of the word "white" bear to all those resembling things.

Again, there have been phenomenalists—and perhaps there are some still—who, while not necessarily denying the existence of properties and classes, have tried to justify the construction of classes of phenomena that share a given phenomenal quality on the basis of some resemblance relation between them that might itself be phenomenally given.

Conceptualists who hold that classifications are mind-dependent, though not, in case of "natural" classifications, arbitrarily imposed on things, will find it appealing to hold that successful comparison of things belongs among the mechanisms of human concept formation.

And even whole-hearted ontological Platonists are likely to admit that there is something in or about concrete things which signals to beings who perceive only causally effective items the fact of their participation in some causally inert Form. Among objects that display various shades of grey (without having any one particular shade "in common," perhaps it is a certain color-resemblance which signals to us their participation in the abstract universal "grey." Again, some sort of resemblance, as the epistemological if not the ontological ground for classification, would seem to serve even the Platonist.

Further, some natural language classifications are almost certainly learned ostensively. One learning theory, perhaps naive but appealing, would have it that Mama brings various red objects to Toddler's attention, each time saying "red," and thus enabling Toddler, after a while, to apply the word "red" to new items that resemble in the relevant respect all of the exemplars previously pointed out by Mama. Latching onto the relevant resemblance relation in such circumstances may be one of the capacities with which humans are born. Even if this view of ostensive learning should turn out to be mistaken, the suggestion is almost irresistible that something like that is going on at the earliest stages of language acquisition. Again, some resemblance-theory of classification (here similarity to exemplars) is central to this view.

While currently favored methods for specifying the truth conditions for atomic sentences, such as "A is white," are good enough for the purposes of logic, they are also fairly uninteresting. That is to be ex-

pected: the simplest predicative sentences are "atomic" to the extent that they reveal no further internal structure that might be of interest to a logician, and almost any method for interpreting them (respecting identities if they occur) will be good enough. Such truth conditions—say, that "Tom is blond" is a true sentence just in case the designatum of "Tom" is in the extension of "blond"—do not, and need not for the purposes of logic, reveal anything about how such a truth-claim might be justified, how the extension of the predicate "blond" might be determined, or how the correct application of such a predicate might be learned. Still, it would seem preferable if truth conditions for such sentences could be specified in terms, such as "resemblance," which express or, if treated abstractly and schematically, at least hint at some procedure that might conceivably be used to *find out* whether such sentences are true.

Even if the implausible claim should be granted that human beings have a special intuitive grasp of participation in Forms or of membership in classes, at least robots surely lack such a faculty of intuition. If robotic devices could be programmed to recognize similarities among objects with respect to whatever features of their input from objects they can be made to respond to, then such programming would seem to have advantages of simplicity and generality and would seem to give promise of building robots capable of feature recognition and classificatory "learning" of a sort which follows principle.

Whereas resemblance theories of classification would offer such obvious advantages, the work of Rudolf Carnap and Nelson Goodman has initiated and justified severe doubt as to whether such theories really work as intended. Notably the well-known difficulties of "companionship" and of "imperfect community" that arise for both phenomenologically and physically based resemblance, and even for resemblance in a given (broad) respect or to a certain degree, are sufficiently severe to warrant doubt whether any such theories can be made to work with sufficient generality.

While classifications based just on pairwise comparisons of things can probably not be effected, we shall make use of a somewhat stronger relation, based on *simultaneous comparison and contrast.* Moreover, the Carnap-Goodman program of constructing quality classes was not explicitly semantical in content. If we think of it instead as a task for specifying truth conditions for sentences of subject-predicate form then, in addition to some relation of similarity, we can also avail ourselves of some relations of *designation* between words and things which make explicit with which things a given item is to be compared, and against which things it is to be contrasted.

Our naive picture of ostensive teaching looks somewhat like this: In teaching the word "white" to Toddler, Mama produces tokens of the word "white" in the presence of some white things and with encouraging demeanor, while also acting discouraging while saying "white" in the presence of some non-white things. Or, perhaps, Mama makes prominent some white things, encourages response, and then hugs Toddler for correct applications of the word (thereby adding them to his positive exemplars) while disapproving of Toddler's mistaken applications (which build up his negative exemplars). Toddler will furnish evidence of having learned the use of "white" intended by Mama if he associates tokens of this word with objects that, in the relevant respect, resemble all of the positively or approvingly designated exemplars, but none of the negatively or disapprovingly designated ones. As Toddler grows up, he might not only learn new bits of language by relying on language already learned, but he might also become sensitive to signals of approbation and disapprobation that are more customary among adults, and his recognition of intended similarities might itself become increasingly sophisticated (as a Platonist would say: he can come to recognize new properties, not just new words.) "Meanings" of simple words, like "white," might be standardized in a language community by comparing sets of positively and negatively designated exemplars used in different teaching situations, and standardized uses might be passed on when Mama uses on her toddler exemplars that are relevantly similar to the ones that were used on her. We allow that the properties with respect to which the comparisons and contrasts proceed might be relative not just to sensory apparatus, but also to interest or features of salience, or to whatever other factors psychologists or anthropologists might find to be relevant. Perhaps properties quite unlike that of being white, and correspondingly a relation of resemblance quite unlike that of color-resemblance, will turn out to be the ones most natural to humans in their infancy (or to robots relying on digitized images).

II. Resemblance, Subject—Predicate Sentences

Formally, let us use a three-place predicate "Res" (short for "resemblance") with the tentative extra-systematic understanding:

(1) Res(x,y,z) just in case, in a given respect, x resembles y but not z.

or, using terminology that an ontological realist would endorse:

(2) Res (x,y,z) if and only if for some quality Q, x has Q and y has Q but z lacks Q

Of course, "Res" is taken to be a primitive predicate, and its explication in terms of quality-possession is meant to be strictly informal and not intended to endorse an ontology with qualities. Also, what shall count as a *quality*, in this informal context, is meant to be open to various interpretations. For the purposes of a phenomenalistic program, perhaps certain features of appearances or "qualia" in the sense of Goodman are wanted; if so, resemblance will proceed in accord with them. For other purposes, qualities of concrete objects, perhaps only dispositionally or indirectly sensible, will be the right ones. In scientific contexts, we might wish to consider qualities which are not at all obvious to the layman but make for simple formulations of scientific laws. If so, comparisons and contrasts which yield the intended judgements of resemblance with respect to such qualities will themselves presuppose scientific training, observation "tainted" by theorizing, and perhaps all manner of apparatus and calculation needed in designating positive and negative exemplars and in formulating and applying criteria for judging that items found in repeated experiments are relevantly similar to ones claimed to have been discovered earlier. In short, whatever qualities a particular inquiry may demand are the ones with respect to which we shall also allow a corresponding relation of resemblance.

However, until further notice, we exclude from consideration universal and empty qualities (which have no negative or no positive exemplars whatever). We shall return to such special cases later. With this understanding, we can define, using our three-place predicate "Res" also a two-place predicate, say "Sim" for "similarity":

(3) *Definition.* Sim(x,y) if and only if for some z, Res(x,y,z).

The relation Res should satisfy the following obvious postulates (and perhaps a few more):

(4) $\sim$Res(x,y,x) (In a given respect, nothing resembles something and not also itself)

(5) Res(x,y,z) $\rightarrow$Res(x,x,z) (If, in a given respect, x resembles y but not z then, in a given respect, x resembles itself but not z)

(6) Res(x,y,z) $\rightarrow$Res(y,x,z) (If in a given respect, x resembles y but not z then, in a given respect, y resembles x but not z).

Anything like the transitivity of resemblance is, of course, not wanted. Nor do we adopt an identity condition such as this one:

(7) $\forall u \forall v[\text{Res}(x,u,v) \leftrightarrow \text{Res}\ (y,u,v)] \rightarrow x = y$ (Roughly: items which resemble all of the same things are identical.) for reasons that were exemplified elsewhere[4].

Since we want to make semantic use of this relation Res in interpreting sentences of subject-predicate form, we must assume that mention of it can be made in the meta-language. Perhaps it is a primitive there, together with some other machinery that might be acceptable to a nominalist and whose precise nature remains to be explored. Perhaps the meta-language expresses instead some deductive principles (such as those of set theory) that are acceptable to ontological realists, and "Res" is defined in that framework in a manner alluded to by the above informal interpretation and more carefully discussed in the appendix.

For semantical purposes, we shall need at least *two* designation relations defined on predicates that are supposed to apply to non-universal and non-empty quality classes. One of them picks out positive exemplars, will be called "positive designation," or "Des$^+$" for short; and the other one will serve to refer to exemplars which illustrate absence of the quality in question, and will be called "negative designation," or "Des$^-$." Since we are not presently concerned with names, let us agree that names shall also refer to the things named by a designation relation which will simply be called "Des." Perhaps names can be assimilated to predicates, but we shall not here be concerned with this problem. To capture the customary uniqueness condition for the designation of names, let us agree that

no name ever designates more than one thing, though we may allow that non-denoting names do not designate anything. If τ is a name and π is a one-place predicate, a truth condition for a sentence having the form $\ulcorner \tau$ is $\pi \urcorner$ will then look like this:

(8) $\ulcorner \tau$ is $\pi \urcorner$ is true just in case for some x, Des(τ, x) and for all y and z, if Des$^+(\pi, y)$ and Des$^-(\pi, z)$ then Res(x,y,z).

Informally: $\ulcorner \tau$ is $\pi \urcorner$ [say, The Taj Mahal is white] is true just in case, given any positive exemplar y of π [something white] and any negative exemplar z [something non-white], there is a respect in which the designatum of "The Taj Mahal" resembles the former but not the latter.

A nominalist might not wish to define such designation relations on predicates (sign-designs or shapes), but rather on tokens of them. While shape-predicates of the meta-language can presumably be taught by similar methods, we assume here tacitly that designation is really defined on shape-tokens whose recognized similarity is built into the notion of designation: part of Toddler's understanding that Mama *designates* objects, perhaps by pointing while uttering tokens of the word "white," includes the understanding that similar tokens are meant to designate similar items. For reasons of simplicity, and since we are not presently concerned with an interpretation of the meta-language, we shall henceforth ignore this complication and just refer to expression-types of the object-language.

There is another difficulty we are going to leave out of account: a positive exemplar of a predicate like "white" serves as such only on occasions at which it is white, and not on occasions when it might have some other color. Accordingly, the (empirical) process of designation should really be relativized to occasions or to times. However, since that complication is usually ignored in other semantic theories as well, we shall do likewise.

Here is how the ostensive teaching of the adjective "white" might proceed: the teacher designates a number of white things (there must be at least one, since we have confined attention to non-empty property-extensions) and does so positively (with a smile, with a hug, or some other sign of approval). She also designates a number of non-white things (there are some, since universal properties are excluded) and does so negatively (with a frown, a slap, or some such indication that those are no-no things). The pupil is supposed to come with the inborn or acquired capacity to recognize, after a while, what is common to the positive exemplars while being absent from the negative ones, and to associate tokens of the word "white" with new items that share exactly what is common to the positive exemplars and absent from the negative ones.

There are some rules of thumb which make for more efficient teaching or serve to avoid mistaken classifications. With regard to "white," (i) The positive exemplars should all be white (but otherwise as diverse as possible); and (ii) the negative exemplars should all be non-white (but as similar as possible to the positive ones). Still, we face versions of the difficulties familiar from Nelson Goodman's discussion of the *Aufbau*:

First, the "companionship difficulty" can arise: suppose that all white things are also bright, so that we could not find a white and non-bright positive exemplar and the pupil might mistakenly come to think that "white" is supposed to apply to all bright things. Then let the teacher follow the rule (iiia) that, whenever an object turns up that shares a certain quality with all of the white exemplars without itself being white (such as a bright and non-white object), let it be added to the negatively designated exemplars. A bright and non-white thing could then share brightness with all of the white exemplars, but it could not *also* fail to share that quality with all of the negative exemplars.

But suppose that, while all white things are bright, there is also no bright non-white thing that could serve as a negative exemplar. Well, in that case all and only the white things are bright, "white" and "bright" will be co-extensional, and we can hardly do better than class-theorists in characterizing quality classes only up to co-extensionality—at least at this first ostensive level of language teaching.

Yet these three rules of thumb are still not good enough because of difficulties of "imperfect community" such as this one: Suppose that "W" stands for "white" and subscripted "Q"'s stand for other qualities, and we have positive exemplars of W like these

(9) WQ_1Q_2 WQ_2Q_3 WQ_3Q_1

and negative non-white ones like these

(10) Q_1 Q_2 Q_3

and an object having these qualities

(11) $Q_1Q_2Q_3$

Then the object (11) will be mistakenly classified as "white" (since, for each pair drawn from the lists (9) and (10) there is a quality that (11) shares with the former but not with the latter), even though there is no one quality that (11) shares with all of the paradigms in (9).

Still, this last-mentioned difficulty can be overcome by requiring that the object (11) shall itself be added to the negatively designated exemplars. Generally, both "companionship" and "imperfect community" are avoided if we replace rule (iiia) by the slightly stronger rule (iii): whenever an object turns up that is similar to all of the white exemplars without itself being white, then let it be added to the negatively designated exemplars.

As the appendix will show with greater precision and generality, this procedure will provably always yield whatever predicate extensions an ontological realist might deem to be the right ones: To any non-universal and non-empty predicate applying on the basis of a quality Q, we can always assign positive and negative exemplars so that (i) all positive ones have Q, (ii) none of the negative ones have Q, (iii) whenever something is similar to all positive exemplars while lacking Q, it is a negative exemplar and, under these constraints, the truth-condition expressed by (8) will be satisfied exactly by the Q-possessing items.

This result is expressed in language a Platonist would endorse; in particular, an interpretation of "Res" in terms of extensions is presupposed. One cannot show that a proposed truth condition meets Platonistic expectations without borrowing Platonistic terminology. But a nominalist, in fixing the intended extension of a predicate π, need presuppose no more than that π positively and negatively designates suitable exemplars while the primitive "Res" is understood. Of course, if Mama, in teaching a word like "white" is careless or perverse enough to pick exemplars at random, there is no telling what Toddler will associate with that word; but such difficulties can arise for any other truth conditions as well.

III. Classifications by Principle and by Enumeration

So far, universal and empty predicate extensions have been left out of account. One could provide for them separately by the convention that a predicate which does not negatively designate anything shall be true of everything (without resorting to Res), and one which fails to enter positive designation shall apply to nothing. However, another approach seems preferable:

In traditional logic, there is a time-honored but currently unfashionable distinction between *classification by principle*, as opposed to *classification by enumeration*. The former is meant to proceed in accord with some rule or some natural property common to all things so classified. The latter proceeds by stipulatively listing the items to be classified. We can provide for predicates whose intended application is either principled or enumerative by laying down a truth condition like this one:

(12) $\ulcorner\tau$ is $\pi\urcorner$ is true if and only if for some x, Des (τ, x) and either:

(a) there is no z such that $\text{Des}^-(\pi, z)$, and $\text{Des}^+(\pi, x)$, or

(b) there is a y such that $\text{Des}^+(\pi, y)$ and there is a z such that $\text{Des}^-(\pi, z)$, and for all y and z, if $\text{Des}^+(\pi, y)$ and $\text{Des}^-(\pi, z)$ then $\text{Res}(x,y,z)$.

The idea is that the first clause, if applicable, signals that the classification intended by the predicate is meant to be an enumerative one. If π is meant to apply to just $x_1, x_2, \ldots$ then those things are positively designated by the predicate, nothing whatever is negatively designated, and $\ulcorner\tau$ is $\pi\urcorner$ will be true just in case τ positively designates one of the x's listed. We can then also get universal or empty extensions, as special cases of enumerative classifications, by letting π positively designate, respectively, everything or nothing.

The second clause, with its condition that π designates negatively, signals that the classification effected by π is meant to be a principled one. It presupposes that π comes with negative exemplars and that it applies by virtue of the principle expressed by Res.

IV. Relational Sentences

Consider next simple sentences formed by applying some non-comparative relation expression to names, such as "Fido bites Daddy." Somehow, the order between the designata of "Fido" and "Daddy" must be taken into account. We *could* allow that "biting" designates *exemplary ordered pairs.* If we did, then the previous discussion could be extended

to relations without any significant changes. However, since we want to please those nominalists who reject the view that (arbitrary) ordered pairs of individuals are always again individuals while refusing to let non-individuals be designated (perhaps on the grounds that items that do not really exist cannot enter into semantic designation relations), let's look for some other approach.

Given utterances of two conflicting relational sentences, like "Tom outweighs Bill" and "Bill outweighs Tom," there is little about *things* in the world that would render one of them true and the other one false: We have the words "Tom," "Bill" and "outweighs"; the persons Tom and Bill; and the sum (if we admit it) (Tom + Bill) which is identical with the sum (Bill + Tom). Just about the only *things* that differ in the two situations are the utterances themselves, or the ordered concrete name-strings: that a "Tom"-before-"Bill" string is currently being realized, rather than a "Bill"-before-Tom" string. Those, however *are* different: if understood as concrete inscriptions (rather than shapes), a specific "Tom"-inscription cannot simultaneously be to the left and to the right of a given "Bill"-inscription (or be uttered both before and after the other's utterance). Unlike ordered pairs or abstract sequences, inscriptions or utterances come with an order which implies that if they exist, their converses don't. Also, inscriptions with built-in order do exist whenever relational sentences are concretized, and hence whenever the truth of such sentence-inscriptions is at issue. For this reason, we take concrete strings of names to be the things in the world which determine the order in which the named items are to be taken in determining which relational sentence is true.

So far, our similarity or resemblance relation was informally explicated in terms of properties. We could still work with those in the present context, if "relational properties," such as "being bitten by Fido" were admitted. However, there would be lots of those "being bitten by Fido," "being bitten by Rex," etc.) and it is not clear how many positive and negative paradigmatic ones among those it would take to teach a new one of the kind.

Instead, let us (extra-systematically) allow for similarity with respect to relations. Suppose Fido bites Fred and Rex bites Max; what *things* are similar by virtue of entering the biting relation? Not ordered pairs: we disallow them. Nor even inscriptional strings: the string "Fido"-before-"Fred" does not enter the biting relation. However this string has a special property: the property that the respective designata of its parts "Fido" and "Fred" enter the biting relation. This property is half semantical (to grasp it, we have to know what "Fido" and "Fred" designate), and half non-semantical: Fido's biting Fred). We could say that the *strings* "Fido"-"Fred" and "Rex"-"Max" are similar by virtue of this property; but that sounds a bit odd. It seems better to say that the respective designata of "Fido"-"Fred" and of "Rex"-"Max" are similar by virtue of entering the biting relation, but say it by building mention of the respective designata into the similarity relation itself. We would say something like: "Fido"-"Fred" bears similarity-of-respective-designata to "Rex"-"Max," which should make it clear that the relation itself obtains between strings of names but by virtue of their designata.

Replace then the former relation Res by Res* (meaning "resemblance-of-respective-designata"), and the former intuitive (and Platonistic) explication (2) by this one

> (13) Res*(τ, ξ, η) if and only if there are a positive integer n, terms $\tau_1, \ldots, \tau_n$, $\xi_1, \ldots, \xi_n$, $\eta_1, \ldots, \eta_n$ and an n-adic relation R such that $\tau = \ulcorner\tau_1 \ldots \tau_n\urcorner$, $\xi = \ulcorner\xi_1 \ldots \xi_n\urcorner$ $\eta = \ulcorner \eta_1 \ldots \eta_n\urcorner$, the respective designata of $\tau_1, \ldots \tau_n$ and those of $\xi_1, \ldots, \xi_n$ enter R, but those of $\eta_1, \ldots, \eta_n$do not.

By "similarity" we shall then mean the analog of "Sim," defined as in (3), but with "Res" replaced by "Res*."

While we have heeded nominalistic preferences in shunning abstract ordered pairs or sequences as objects and in selecting our resemblance relation, there is no present need to insist formally that concatenates of terms must be inscriptions.

While one could extend the notions of positive and negative designation from one-place predicates to relation expressions, it seems more intuitive to associate with the latter (and hence also with the former as special cases) atomic sentences containing them which are either *true by stipulation* (perhaps better: true by ostensive teaching) and *false by stipulation*. Those positive and negative basic truths serve as intended if they are paradigmatic of a relation's obtaining or not obtaining. Accordingly, we assume that every n-place predicate (whose extension is intended to be non-empty, non-universal, and determined in a principled way) enters into a

relation of positive association and of negative association with atomic sentences containing that predicate which are taken to be, respectively, true or false by stipulation.

Using a new notion—say "True*"—we can then replace and generalize the former truth condition (8) as follows:

(14) A sentence of the form $\ulcorner \pi\tau \urcorner$ (where $\tau = \ulcorner \tau_0 \ldots \tau_{n-1} \urcorner$) is True* just in case for every sentence $\ulcorner \pi\xi \urcorner$ (where $\xi = \ulcorner \xi_0 \ldots \xi_{n-1} \urcorner$) that is positively associated with π and every sentence $\ulcorner \pi\eta \urcorner$ (where $\eta = \ulcorner \eta_0 \ldots \eta_{n-1} \urcorner$) that is negatively associated with π, Res*(τ, ξ, η).

Informally (and Platonistically): a relational sentence of the form $\ulcorner \pi\tau_0 \ldots \tau_{n-1} \urcorner$ is true just in case for every sentence of like form which is true by stipulation and every such sentence that is false by stipulation there is a relation which the respective designata of τ and of the terms of the former sentence enter but the designata of terms of the latter sentence do not.

With respect to the relation expressed by a predicate π, let us assume (1) that the relation is neither universal nor empty, (2) that the designata of terms in π-sentences which are true by stipulation always enter the relation, while (3) the designata of terms in π-sentences that are false by stipulation never do, and also (4) that, whenever the designata of a term-string do not enter the relation while being similar to the term-strings of all π-sentences that are true by stipulation, then the corresponding π-sentence is declared to be false by stipulation. Then the True* sentences of that form will be exactly those that are true by usual truth conditions (see appendix).

The former truth condition (12), which allowed for classifications both by principle and by enumeration, is readily extended, in the obvious way, to relational sentences.

V. Typical Exemplars and Vague Predicates

For the sake of simplicity, let us ignore relation-expressions in the sequel and revert to our former mention of positive and negative exemplars of (monadic) predicates.

We have said little, so far, regarding the intended intuitive notion of similarity at issue. There are various such notions: There is (1) that of having something like a quale "in common," (2) that of indistinguishability of two uniform qualitative items from all third such items, (3) that of indistinguishability of two items by direct comparison, (4) that of just bare distinguishability, and (5) that of falling within a quality range.

While our truth condition is neutral between all of these senses, it would seem best to think of positively designated items as *typical* exemplars and of those objects as similar to them which lie in a (qualitative) *neighborhood* of the typical ones. Thus, in teaching a word like "grey," we would pick positive exemplars which display a typical shade of grey (or, since there might be missing shades of color, perhaps several shades of grey which come closest to being typical) and call those objects "grey" whose color falls within the common neighborhood of the positive exemplars. While most qualitative neighborhoods will not be uniquely delimited (in as much as most predicates of natural discourse are vague), application of our semantics in a given context presupposes that exactly those objects resemble the typical exemplars of a given predicate, such as "grey," to which, due to properties in the corresponding neighborhood of those exemplified, the predicate would be said to apply. Such qualitative neighborhoods might be multi-dimensional. E.g., the neighborhood of a typical green shade might include bluish-green, yellowish-green, saturated and faded green, dark as well as light green. Items negatively designated by "green" should accordingly comprise ones which fall outside the neighborhood of green in all of these respects.

The truth conditions for generic sentences, such as "stars twinkle," do not seem to depend on quantifications over all things or all stars: not all stars twinkle; in fact, most stars don't; it does not seem enough that some of them do so some of the time. The things that first come to mind as one thinks of stars, the "typical stars," are wont to twinkle; but the interpretation of "typical" has also remained elusive so far. However, if the typical stars are just those that served as paradigms in our own experience of learning the word "star," if they are just the positive exemplars of the predicate "being a star," then many such generic sentences can readily be interpreted in the framework of a semantics which assigns to predicates typical exemplars.

To the extent that generic sentences can be taken to mean the same thing as general assertions about typical representatives of the genus—e.g., "birds are winged" as meaning "all typical birds are winged," or "grey things are not colored" meaning "all typi-

cally grey things lack color"—such sentences can apparently be accommodated if we assume that the positive exemplars of the generic term are indeed typical and have been chosen with enough diversity so that exactly the properties of typical instances are common to them all. Basic phrases of the form "a typical so-and-so" can thus be interpreted in our proposed semantics as being positively designated by "so-and-so."

Let us turn next to *vague* predicates. Among them we focus first on what seem to us the basic kinds, namely ones which are "conjunctive," such as "hot" in "hot pan," by virtue of implying "being hot and being a pan," in teaching a vague predicate which has a natural opposite, like "hot," Mama would choose positive exemplars which are typically hot in the child's experience; say an open flame, but not the sun. And she would select negative exemplars that are typically cold; perhaps ice cubes, but not liquid hydrogen. With added experience regarding the use of "hot," the pupil may learn that the temperature of the sun is also reckoned to be in the neighborhood of the temperature of an open flame, for the purpose of applying "hot." From our point of view, vague terms reflect a vaguely applied notion of similarity in assessing the conditions of their application. And vague terms can be stretched along with the notion of a common "quality" which accompanies the relation of similarity.

Next, we would like to accommodate *comparatives* corresponding to vague predicates, like "hotter." To this end, we shall need to express some sort of distance between typical exemplars in qualitative neighborhoods.

Among various options for expressing comparatives, some (such as comparative resemblance) do not seem to yield our present relation Res as a special case while remaining asymmetric. Instead, let us suppose that vague predicates (say, "hot") that correspond to comparatives (like "hotter") are learned first, and something like the expression "*a* is more so than *b*" is learned subsequently. We may then suppose that not only comparisons (like those expressed by "Res") *in some respects or other*, but also comparisons *in a given respect* π are available, where π is a predicate (not a property) whose use has been previously learned.

For the purpose of interpreting "*A* is hotter than *B*," and given that fire is paradigmatically hot, employment of an expression such as "with respect to "hot," *A* is more like fire than is *B*" comes to mind. However, if we think of positive exemplars as being typical, rather than extreme, this will not do: molten iron resembles a flame more closely than does the sun; yet the former is not hotter than the latter. Instead, we need a primitive, say "Unl," entering into formulas like "Unl("hot," sun, molten iron, ice)" expressing that, with respect to the predicate "hot," the sun, more than molten iron, is *unlike* ice (a negative exemplar of "hot").

A Platonist who avails himself of something like metric neighborhood spaces might endorse the following informal explication of the semantic primitive "Unl"

> (15) Unl (π, *A*, *B*, *Z*) just in case, with respect to the neighborhood system assigned to π the distance between *A* and *Z* is greater than the distance between *B* and *Z*.

In case of a comparative like "hotter" which comes from the "conjunctive" predicate "hot," it seems to accord with our analytic expectations that, whenever one thing is hotter than another, both are still hot. Thus, it seems incorrect to say that ice is hotter than liquid hydrogen, inasmuch as both are cold (though the former has a higher temperature than the latter, and the latter is colder than the former). Not all comparatives are like that; but let us consider the others in a moment.

On a semantic level, by at least one recursive step higher than the one on which truth conditions for the predicate π (or perhaps some arbitrary formula $\varnothing$) have been defined, one can then say under what circumstances it shall be true that τ, more than ξ, is π:

> (16) ⌜ More-than[$\pi\tau$, $\pi\xi$] ⌝ is true just in case ⌜ $\pi\tau$ ⌝ and ⌜ $\pi\xi$ ⌝ are both true, and for all *x*, *y*, and *z*, if Des (τ,*x*), Des (ξ,*y*) and Des (π,*z*) then Unl(π, *x*,*y*,*z*).

E.g., τ, more than ξ is hot just in case both τ and ξ are hot and, with respect to "hot," the designatum of τ, more than the designatum of ξ, is unlike all negative exemplars of "hot."

Now let us turn to vague predicates which are not "conjunctive" and whose applicability seems to be relative to a reference class, such as "tall" in "tall midget," or "small" in "small elephant." Equipped with an interpretation of "typical" and with comparatives, it seems that we can interpret "*A* is a small elephant" to mean (among people who know about

elephants) "*A* is an elephant and typical elephants are larger than *A*," with the implication we want to bestow on "larger" that both typical elephants (i.e., positive exemplars of "elephant") and *A* are large. Generally, in the slightly paradoxically sounding occasions when we want to assert something of the form "*A* is a ø so-and-so," while typical so-and-so's are known to be non-ø, the assertion seems to imply "all typical so-and-so's are less ø than is *A*," with the understanding that "being less-ø" implies "not being ø," just as "being hotter" implies "being hot."

Suppose I tell you: "Uli is a frail Hornussen player," and you have no idea about what typical Hornussen players are like. In that case you would probably feel misled if Uli turned out to be a fairly robust and stout fellow. Yet, once you came to know that typical players of that game are strong men, you would take me to have meant that typical Hornussen players are more robust than Uli. If so, the more basic meaning of "being a frail so-and-so" appears to be the one which implies "being frail," while its variant presupposes factual knowledge regarding so-and-so's, and had perhaps best be regarded as a linguistic liberty sanctioned in the presence of common background knowledge.

Just as "being a tall midget," though slightly strange, is a phrase to whose intended meaning we can readily adjust, so also to say of two midgets *A* and *B* "*A* is taller than *B*" might imply in isolation that both are tall, but need not be so taken, once it is clear that we are talking about midgets. Whereas it would seem more appropriate to say, in such a case, "*B* is shorter than *A*"—with the expected implication that both *A* and *B* are short—the reverse, slightly paradoxical assertion would again seem to be one we tolerate, like ever so many other slight inaccuracies, in the presence of shared knowledge to the effect that *A* and *B* are midgets.

Thus, a semantics which makes use not only of extensions of predicates (conceived as comprising those items to which the predicate properly applies), but also of positive exemplars (conceived as items to which the predicate applies paradigmatically) will have the technical tools for interpreting phrases of the form "a typical so-and-so," and thereby also machinery for an improved interpretation of generic sentences, like "birds fly," and of non-conjunctive predicates.

The University of Rochester

Received September 5, 1989

APPENDIX: A PROOF

We show here that, under the three assumptions informally mentioned in section II (but adjusted here to the general notions discussed in section IV), the truth condition (13) in terms of resemblance coincides with customary truth conditions:

Assume that

(1) π is an n-place predicate

(2) Ext(π) is the extension of π,

(3) Convention: $\tau = \ulcorner \tau_0 \ldots \tau_{n-1} \urcorner$, $\xi = \ulcorner \xi_0 \ldots \xi_{n-1} \urcorner$, $\eta = \ulcorner \eta_0, \ldots, \eta_{n-1} \urcorner$ [i.e., τ, ξ, and η abbreviate n-adic strings of terms],

(4) If $\mathrm{Des}^+(\pi, \psi)$ then, for n names $\xi_0 \ldots \xi_{n-1}$, $\psi = \ulcorner \pi\xi_0 \ldots \xi_{n-1} \urcorner$ [i.e., π positively designates only atomic sentences prefixed by π: those that are true by stipulation],

(5) If $\mathrm{Des}^-(\pi, \chi)$ then, for n names $\eta_0, \ldots, \eta_{n-1}$, $\chi = \ulcorner \pi\eta_0 \ldots \eta_{n-1} \urcorner$ [i.e., π negatively designates only atomic sentences prefixed by π: those that are false by stipulation],

(6) $\mathrm{Res}^*(\tau, \xi, \eta)$ iff for some, positive integer n, some n-adic relation R, and 3 n-adic sequences of terms, τ, ξ, and η, $\langle \mathrm{Des}(\tau_i) \rangle_{i<n} \in R$ and $\langle \mathrm{Des}(\xi_i) \rangle_{i<n} \in R$ but not: $\langle \mathrm{Des}(\eta_i) \rangle_{i<n} \in R$ [this is interpretation (13) of Res*: the respective designata of τ resemble those of ξ but not those of η just in case τ, ξ, and η, are strings of terms such that, for some relation R, the respective designata of τ and ξ enter R, but those of η do not],

(7) A sentence of the form $\ulcorner \pi\tau \urcorner$ is True* iff for all sentences $\ulcorner \pi\xi \urcorner$, $\ulcorner \pi\eta \urcorner$, if $Des^+(\pi, \ulcorner \pi\xi \urcorner)$ and $Des^-(\pi, \ulcorner \pi\eta \urcorner)$ then $Res^*(\tau, \xi\eta)$ [this is the truth condition (14)],

(8) for some sentence ψ, $Des^+(\pi, \psi)$ [this rules out predicates which never apply],

(9) for some sentence χ, $Des^-(\pi, \chi)$ [this rules out predicates which apply to everything],

(10) Whenever $Des^+(\pi, \ulcorner \pi\xi \urcorner)$, the sequence of designata of the ξ's $<Des(\xi_i)>_{i<n} \in Ext(\pi)$[roughly, sentences that are true by stipulation are really true; this corresponds to former rule (i)],

(11) Whenever $Des^-(\pi, \ulcorner \pi\eta \urcorner)$, the sequence of designata of the η's $<Des(\eta_i)>_{i<n}$ is not in $Ext(\pi)$ [sentences that are false by stipulation are really false; this is former rule (ii)], and

(12) whenever $<Des(\tau_i)>_{i<n}$ is not in $Ext(\pi)$ and for all ξ such that $Des^+(\pi, \ulcorner \pi\xi \urcorner)$ there is an η such that $Res^*(\tau, \xi, \eta)$, then $Des^-(\pi, \ulcorner \pi\tau \urcorner)$ [whenever the designata of τ_i are not in the extension of π while they are similar to all of the respective designata of π-sentences that are true by definition, then $\ulcorner \pi\tau \urcorner$ is false by definition; that corresponds to former rule (iii)].

Then $\ulcorner \pi\tau \urcorner$ is true (in the conventional sense) if and only if $\ulcorner \pi\tau \urcorner$ is True*.

Proof: Sufficiency: assume that $\ulcorner \pi\tau \urcorner$ is true, so that $<Des(\tau_i)>_{i<n} \in Ext(\pi)$. Using assumption (7), show $\ulcorner \pi\tau \urcorner$ is True*. Assume $Des^+(\pi, \ulcorner \pi\xi \urcorner)$ and $Des^-(\pi, \ulcorner \pi\eta \urcorner)$. By (4) and (10), $<Des(\xi_i)>_{i<n} \in Ext(\pi)$ and by (5) and (11), $<Des(\eta_i)>_{i<n}$ is not in $Ext(\pi)$. Hence, by (6), $Res^*(\tau, \xi, \eta)$. *Necessity:* assume that $\ulcorner \pi\tau \urcorner$ is True*. Show: $<Des(\tau_i)>_{i<n} \in Ext(\pi)$. Suppose not. Claim: for all ξ such that $Des^+(\pi, \ulcorner \pi\xi \urcorner)$ there is an η such that $Res^*(\tau, \xi, \eta)$. Assume $Des^+(\pi, \ulcorner \pi\xi \urcorner)$. By (4) and (10), $<Des(\xi_i)>_{i<n} \in Ext(\pi)$. By (9), (5) and (11), there is an η, such that $<Des(\eta_i)>_{i<n}$ is not in $Ext(\pi)$. Hence, by (6), $Res^*(\tau, \xi, \eta)$, establishing the claim. But then, by (12), $Des^-(\pi, \ulcorner \pi\eta \urcorner)$. By (8) and (5), for some ξ, $Des^+(\pi, \ulcorner \pi\xi \urcorner)$. But then, since $\ulcorner \pi\tau \urcorner$ is True* and by (7), $Res^*(\tau, \xi, \pi)$, which is impossible by postulate (4). Q.E.D.

Remark 1. Clearly, if *all* true π-sentences are true by stipulation and *all* false ones are false by stipulation, the assumptions (10)-(12) are satisfied. So, there is at least one assignment of positive and negative paradigms due to which the True* sentences are exactly the intended ones.

Remark 2. The monadic case, discussed in sections II, III, and IV is clearly a special case of the above: whenever $Des^+(\pi, \ulcorner \pi\xi_o \urcorner)$, regard $Des(\xi_o)$ as a positive exemplar of π; and when $Des^-(\pi, \ulcorner \pi\eta_o \urcorner)$, treat $Des(\eta_o)$ as a negative exemplar.

NOTES

1. Carnap, R., *The Logical Structure of the World & Pseudoproblems in Philosophy*, translated by R. A. George (Berkeley and Los Angeles: University of California Press, 1967).

2. Goodman, N., *The Structure of Appearance*, 3rd ed. (Dordrecht: D. Reidel Publishing Company, 1977).

3. Eberle, R. A., "Appendix: Predication and Regresses; Classification by Resemblance," in *Language, Truth, and Preciation*, by Madhabendranath Mitra (New Delmi: New Statesman Publishing, 1988), pp. 160-81. A brief and preliminary version of the present proposal (not yet capable of accommodating relation expressions and not yet explicit about the paradigmatic role of positively designated items) is surrounded there by discussions of other related topics.

4. Eberle, R. A., *Nominalistic Systems* (Dordrecht: D. Reidel Publishing Company). A counter-example to (7) occurs on p. 169.

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 27, Number 4, October 1990

A READING OF AQUINAS'S FIVE WAYS

Robert J. Fogelin

As they appear in the *Summa Theologica*, Thomas Aquinas's so-called Five Ways are bracketed by two objections and two replies to these objections. Now in commenting on the five ways, many twentieth-century writers on Aquinas, including Kenny,[1] Copleston,[2] and Gilson,[3] make no reference to these objections and the replies to them. This is surely odd, for throughout the *Summa Theologica* objections and replies provide the framework within which Aquinas's unfolds his ideas, and they thus provide basic keys for interpreting the text. With this in mind, and in contrast with the writers just mentioned, I propose to examine Aquinas's Five Ways as answers to the objections that precede them and as grounds for the replies that follow them. As we shall see, this approach is not without difficulties, but, in any case, it yields a reading of the Five Ways that is different, indeed radically different, from the interpretations that have now become standard.

The first objection that Aquinas raises concerns the problem of evil. He asks how the existence of an infinitely good (*bonum infinitum*) being is compatible with the presence of evil in the world. The problem, as Aquinas initially states it, is how such an infinitely good being could, so to speak, *leave room* for bad things. He puts it this way:

> For if, of two mutually exclusive things, one were to exist without limit, the other would cease to exist. But by the word 'God' is implied some limitless good (bonum infinitum). If God then existed, nobody would ever encounter evil. But evil is encountered in the world. God therefore does not exist. (Ia, 2, 3)[4]

Read uncharitably as an argument concerning infinite *magnitudes*, the reasoning would, of course, be fallacious. The existence of infinitely many even numbers, for example, is compatible with the existence of infinitely many odd ones, even though being even is incompatible with being odd. Or again, the existence of an infinite region of luminosity ending at South Bend, is compatible with the existence of an infinite expanse of darkness beginning there. But surely Aquinas had in mind infinite goodness combined with the other traditional attributes of God, and he is asking how an infinitely good being could *allow* the existence of evil. Indeed, when we turn to his response to the objection, we see that Aquinas attempts to answer question this precisely.

> As Augustine says, *Since God is supremely good, he would not permit any evil at all in his works, unless he were sufficiently almighty and good to bring good even from evil.* It is therefore a mark of the limitless goodness of God that he permits evils to exist, and draws from them good. (Ia, 2, 3)

Now my suggestion that we interpret the Five Ways in the light of the objections and responses that enclose them gains little support from this first objection and the response to it. Although some of Aquinas's proofs bear upon this objection (in particular, the fourth and fifth proofs), I do not think that any of them addresses it directly. Even so, this objection and the one to follow share a common feature: They are both challenges to the existence of a traditional Christian God. The first objection provides a reason for saying that God *does not* exist. The second objection, as we shall see, draws the weaker conclusion that *there is no need to suppose* that a God exists. In his response to the first objection, Aquinas clearly attempts to vindicate antecedently accepted Christian doctrine in the face a specific challenge. The leading idea of this essay is that the Five Ways are intended to supply a vindication of Christian doctrine specifically in response to the challenge presented in the second objection.

Objection 2 reads as follows:

> [I]f a few causes fully account for some effect, one does not seek more. Now it seems that everything we observe in this world can be fully accounted for by other causes,

> without assuming a God. Thus natural effects are explained by natural causes, and contrived effects [i.e., purposeful acts] by human reasoning and will. There is therefore no need to suppose that God exists. (Ia, 2, 3)

For the understanding of the natural world, God, this objection tells us, is, as Laplace was supposed to have said to Napoleon, an unnecessary hypothesis.

The form of this objection specifies the form that Aquinas's response must take. He will have to show:

1. Nature is not fully explicable in terms of *natural* causes, and

2. The purpose found in the universe cannot be fully explained by an appeal to *human* reason and *human* will.

With these threats removed, Aquinas will have vindicated the traditional Christian belief in God the creator and sustainer of the world specifically against the charge that this belief is unnecessary for understanding the world. This, I suggest, is precisely what the Five Ways are intended to show. Despite the general language ("There are five ways to prove [*probari*] that God exists"), the point of these *proofs* might better be expressed in these words: *There are five ways of showing that appeals to natural principles and appeals to human reason and human will do not wholly explain natural phenomena. Thus for a complete explication of natural phenomena, these natural principles must be supplemented by an appeal beyond the natural realm.*

Let me say at once that some textual evidence creates a presumption against this reading. At Ia, 2, 2, the article immediately preceding the presentation of the Five Ways, Aquinas speaks explicitly of "demonstrating from effects that God exists," and says quite explicitly:

> God's effects ... can serve to demonstrate that God exists, even though they cannot help us to know him comprehensively for what he is.

On the other side, in the presentation of the Five Ways, Aquinas nowhere refers to his arguments as demonstrations, instead, he speaks of them as proofs. Now to prove something can mean to put it to the test and it is in this sense that I think that Aquinas is attempting to prove the existence of God. But I do not want too much to turn upon the difference between a demonstration and a proof, for the question at issue is what form Aquinas's arguments actually take. Accepting the immediate context of the objections and replies as controlling, I shall argue, despite earlier suggestions at Ia, 2, 2 to the contrary, that Aquinas is not offering demonstrations of the existence of God based upon natural principles,[5] but is, instead, arguing that a recognition of the explanatory *inadequacies* of natural principles forces us to go beyond them.[6]

The Fifth Way

The pattern of reasoning that I am attributing to Aquinas emerges clearly in the Fifth Way, where Aquinas explicitly responds to the claim that human reason and human will can be taken as the cause of all purposeful acts. There he tells us:

> An orderedness of actions to an end is observed in all bodies obeying natural laws, even when they lack awareness. For their behaviour hardly ever varies, and will practically always turn out well; which shows that they truly tend to a goal, and do not merely hit it by accident. (Ia, 2, 3)

Provided that we are willing to join Aquinas in ascribing final causes to the activities of physical objects, we are thus presented with a vast number of examples of purposive events that cannot be attributed to the operations of human reason and will.

This much, however, does not show that something *beyond* nature is needed to account for the purposiveness in the world. We might think of final causes as themselves natural principles needing no external support; that is, the objector could argue that a *natural* teleology is adequate to explain all the purposive events found in the world, and therefore, with respect to these purposive events, the hypothesis of a divine cause is unnecessary. Against this claim, Aquinas responds with breathtaking brevity:

> Nothing however that lacks awareness tends to a goal, except under the direction of someone with awareness and with understanding; the arrow, for example, requires an archer. Everything in nature, therefore, is directed to its goal by someone with understanding, and this we call God'. (Ia, 2, 3)

As a response to a natural teleology this is, of course, baldly question-begging. We might better say that Aquinas is gesturing toward an argument rather than giving one. Yet the pattern of the reasoning is clear: Aquinas is arguing that appeals to merely human teleology or to merely natural teleology are inade-

quate for the explanation of what we observe in nature, and therefore these principles demand for their completion an appeal to God.

Furthermore, read as a freestanding demonstration of the existence of God, the argument is just awful. Even if we grant that the purposiveness in the world cannot be explained by appeals to natural teleology or human teleology, and further grant that it must be explained by an appeal to a supernatural intelligence, it still does not follow that we must postulate the existence of anything like a traditional Christian deity to account for it. To get *that* result, a detailed, full-blown argument from design is needed that eliminates alternative hypotheses to the postulation of a being possessing the unity, perfection, providence, etc., of the traditional Christian God. Since Aquinas provides no such argument, the Fifth Way, read as a teleological proof of God's existence, is simply a failure. Yet if we read it as an attempt to vindicate the claim of God's existence (to *prove* it in this sense) against the counter-claims of natural science, the rhetorical situation changes in an important way. The dialectical development unfolds as follows:

1. We begin with the presumption that the world is the product of a Christian God's creation.

2. The natural scientist tells us that the world can be fully explained on natural principles; therefore, this presumption is idle and can be set aside, at least when we are doing natural science.

3. It is then argued that the world (here with respect to its teleological features) cannot be fully explained by natural principles, therefore, the presumption in 1 is restored.

Read in this manner, the Fifth Way is not a sketchy, incomplete (bad) argument from design intended to establish the existence of a deity with certain determinate features. On the contrary, a belief in the existence of a deity with certain determinate features is antecedently given, and the Fifth Way is an attempt to defend the belief in such a being against a specific challenge.

The Fourth Way

Continuing in reverse order, Aquinas's Fourth Way concerns "gradations found in things:"

> Some things are found to be more good, more true, more noble, and so on, and other things less. But such comparative terms describe varying degrees of approximation to a superlative; for example, things are hotter and hotter the nearer they approach what is hottest. Something therefore is the truest and best and most noble of things, and hence the most fully in being; for Aristotle says that the truest things are the things most fully in being. Now *when many things possess some property in common, the one most fully possessing it causes it in the others.* ... There is something therefore which causes in all other things their being, their goodness, and whatever other perfection they have. And this we call God." (Ia, 2, 3)

As a first approximation, the argument seems to run as follows:

1. We observe in nature that some things are better, truer, or more noble than others.

2. But such ascriptions presuppose a top member of each scale (the superlative), which serves as the standard by which such gradations are established.

Therefore:

> The best, truest, most noble things must exist.

Of course, stated this way, the argument has no tendency to prove the existence of a deity. With respect to beauty, for example, some corporeal entity, perhaps Helen, might occupy the top rung. Something more is needed to force us beyond the natural order. As Kenny puts it:

> It is at this point that St. Thomas needs to appeal implicitly to Plato to fill the gap; for on Plato's view, to be more or less *F* precisely is to participate more or less fully in the Idea of *F* which is the most *F* thing, the one and only thing which is fully *F*. (p. 81)

What form should this appeal to Plato take? Here it is tempting to go back to Plato's writings and examine his (various) reasons for introducing forms, and then ask how the doctrine of forms, so introduced, can be serviceable for Aquinas's purposes. Taken this way, the Fourth Way will swim or, more likely, sink relative to the plausibility of Plato's theory of forms. This approach leads to a deep and complicated inquiry, since many of Aquinas's Aristotelian commitments place him in opposition to these Platonic doctrines.[7] But perhaps Aquinas's appeal to Platonism has a more general form, and goes something like this: there can be no natural explanation of the ranking of natural things, for all such rankings

presuppose the existence of an antecedent standard that gives each of these natural objects a place in the ranking. So even if Helen was, *de facto*, the most beautiful woman, we recognize that even she might have been more beautiful—with a face that could, let's say, launch 1500 ships—and thus her place at the top of the natural order can be grasped only against the background of an ideal ordering. What is needed here is an argument showing that no natural object can, *in principle*, stand at the top of a hierarchy of values, thus determining it. Aquinas presents no such argument, but Aristotle waves a hand in this direction in his early, Platonic-sounding (largely lost) dialogue *On Philosophy*. I shall quote the surviving argument in full:

> Where there is a better, there is a best; now among existing things one is better than another; therefore there is a best which must be divine.[8]

Admittedly, this is rather enthymematic, for no justification is given for the claim that the best "must be divine." In any case, my suggestion is that in the Fourth Way, Aquinas, like Aristotle, is arguing that certain things found in the natural order (in this case, rankings according to value) cannot be explained on natural principles alone. Contrary to Objection 2, this is another respect in which nature cannot take care of itself; thus the *prior* presumption in favor of a transcendental source of value is preserved.

The Three Cosmological Ways

If we take the denial of the possibility of an infinite regress as the signature of a cosmological argument, then Aquinas's first three Ways count as cosmological arguments. The first concerns *motion*, the second *efficient causality*, the third *necessity and contingency*. In each case the arguments are presented against the background of Aquinas's own complex, sometimes obscure, metaphysical position. For these reasons, these arguments do not have the accessibility suggested by their standard appearance in introductory philosophy anthologies.

Since the argument in the Third Way refers back to the argument in the Second, I shall reverse my reverse order and examine the Second Way before the Third.

The Second Way

Sharply edited, the Second Way reads as follows:

> The second way is based on the nature of [efficient] causation.[9] In the observable world causes are found to be ordered in series; we never observe, nor ever could, something causing itself, for this would mean it preceded itself, and this is not possible. Such a series of causes must however stop somewhere ... One is therefore forced to suppose some first cause, to which everyone gives the name 'God.' (Ia, 2, 3)

In presenting the Second Way, I have here deleted the argument intended to show that an infinite regress of efficient causes is impossible - i.e., I have left out the core argument of the Second Way. I've done this to bring the point of the argument into prominence. What the passage seems to say is that the impossibility of an infinite regress of efficient causes demands that sometime *back* in this series of efficient causes, a first (uncaused) efficient cause must exist. Such a reading is actually forced on us by the second sentence in the passage, which clearly indicates that all efficient causes in the observable world are temporally prior to their effects.[10] Borrowing the term from Copleston, I will say that treating efficient causation in this way yields the *horizontal* interpretation of the Second Way. (p. 123)

However natural (and I think forced) this horizontal reading of the text may seem, it is now generally rejected in favor of a *vertical* reading. Copleston describes the difference between these two interpretations in these words:

> [W]hen Aquinas talks about an 'order' of efficient causes he is not thinking of a series stretching back into the past, but of a hierarchy of causes, in which a subordinate member is here and now dependent on the causal activity of a higher member. (p. 122)

Thus:

> We have to imagine, not a lineal or horizontal series, so to speak, but a vertical hierarchy, in which a lower member depends here and now on the present causal activity of the member above it. (p. 123)

A similar vertical interpretation of the Second Way is found in, for example, works of Kenny, Gilson, Garrigou-Lagrange, and others.

Since it is contrary to the simplest reading of the text, why should the vertical interpretation of the Second Way be preferred to the horizontal interpretation? The main reason that commentators reject the horizontal interpretation is that it seems to con-

flict with other texts in Aquinas's writings. Copleston describes the situation this way:

> [T]hough as a Christian theologian [Aquinas] was convinced that the world was not created from eternity he stoutly maintained that philosophers had never succeeded in showing that creation from eternity is impossible....That is to say, no philosopher had ever succeeded in showing the impossibility of a series of events without a first assignable member. (p. 57)[11]

Although Copleston cites no passage to support this claim, his wording on page 122 of his Aquinas suggests that he almost certainly has in mind the following difficult passage cited explicitly by Kenny:

> It is not impossible to go on for ever *per accidens* in a series of efficient causes...as a smith may act by using many different hammers, *per accidens*, if one after the other is broken. For it is not essential for any particular hammer to act after the action of another, and it is likewise not essential for any particular man, *qua* begetter, to be begotten by another man; for he begets *qua* man, and not *qua* son of another man....Hence it is not impossible to go on for ever in the series of men begetting men; but such a thing would be impossible if the generation of one man depended on another and on an element, and on the sun, and so on to infinity. (Ia, 46, 2, 7)[12]

Commenting on this passage, Kenny tell us:

> The series of causes in the Second Way...does not stretch backwards in time, but stretches into the heavens simultaneously. It is this series which must come to an end with God. (p. 42)

Before commenting on Ia, 46, 2 , 7 directly, it is important to be clear about the argumentative situation. Both Copleston and Kenny seem to be reasoning in the following way:

1. In the Second Way, Aquinas is attempting to demonstrate the existence of God.

2. A key move in the argument is that a certain kind of natural infinite regress is impossible.

3. But Aquinas (elsewhere) acknowledges that no philosopher has ever succeeded in showing the impossibility of a series of events without a first assignable *temporal* member.

Therefore: Aquinas was not arguing that the series of *temporally antecedent* efficient causes must terminate in a first cause, but, instead, he was arguing that the system of *simultaneous* efficient causes must terminate in a first cause.

What does Ia, 46, 2, 7 actually say? First of all, it does not speak *directly* about a contrast between temporally prior and simultaneous efficient causation. At most, it cites examples that suggest this contrast. I say *at most* it suggests this, for although the series of begetters (and the series of hammers) does stretch back into the past, thus supplying a clear example of temporal antecedence, the allusion to "an element and the sun" does not carry an obvious reference to simultaneity. More to the point, the central theme of the passage is not about temporal antecedence as opposed to simultaneity, but about something else: a contrast between efficient causes *per se* and efficient causes *per accidens*. This is not evident in Kenny's citation of the passage, for he has simply deleted all references to efficient causes *per se*. Here is the entire passage with the material Kenny has deleted given in bold print.

> **An infinite series of efficient causes essentially subordinate to one another is impossible, that is causes that are *per se* required for the effect, as when a stone is moved by a stick, a stick by a hand, and so forth: such a series cannot be prolonged indefinitely.** [But] it is not impossible to go on forever *per accidens* in a series of efficient causes, **as when they are all ranged under causal heading and how many there are is quite incidental,** as a smith may act by using many different hammers, *per accidens*, if one after the other is broken. For it is not essential for any particular hammer to act after the action of another, and it is likewise not essential for any particular man, *qua* begetter, to be begotten by another man; for he begets *qua* man, and not *qua* son of another man. **For all men in begetting hold the same rank in the order of efficient causes, namely that of being a particular parent.** Hence it is not impossible to go on forever in the series of men begetting men; but such a thing would be impossible if the generation of one man depended on another and on an element, and on the sun, and so on to infinity.[13]

As far as I can see, this passage makes no reference to, and has no implications for, a contrast between simultaneous (vertical) efficient causation and temporally prior (horizontal) efficient causation. The passage concerns the contrast between causes *per se* and causes *per accidens*. I shall try to give a relatively simple explanation of this distinction.

Using Aquinas's examples as our guide, we might

explain the distinction between causes *per se* and causes *per accidens* in the following way: sometimes we pick out (refer to, name) a cause by using language that indicates *how* it acts as a cause. That is, sometimes in indicating a cause we use descriptive terms that are nomologically relevant. If asked who caused a disturbance, I might reply that the man waving a political sign caused the disturbance, thereby indicating (but not directly saying) how he caused it. Here (without obvious qualifications needed) I shall say that I have identified the cause *per se*. I might also merely pick out the cause in a nomologically irrelevant way by saying, perhaps, that the man wearing the grey shirt caused the disturbance. Here the cause has been identified, but only *per accidens*.[14]

Reading Aquinas this way make him into a proto-Davidson; but he could be in worse company.[15] Anyway, the reading makes good sense of Aquinas's examples. It is not in virtue of being begotten that a father can beget a child, even though if he had not been begotten, he would not be available as a begetter. As a member of the series of begetters and begotten, he is only *per accidens* the cause of the child. Viewed in this nomologically arbitrary way, an infinite regress is not repugnant to reason.[16] The situation is different with efficient causality *per se* where we are concerned with *dependence* between the cause and effect. It is offensive to reason (many, including Aquinas, think) to have a chain of dependence that lacks an ultimate mooring. We might call this the *brute cosmological instinct*. Looked at this way, there is nothing offensive to reason in a chain of causes *per accidens* going back (one damn thing before another) infinitely far. It is offensive to reason for a series of causes *per se*, a series of dependencies, to do this, since it leaves the brute ontological instinct unsatisfied.

To go back to the Second Way, Ia, 46, 2, 7 so interpreted, provides a guide for elucidating the core argument of the Second Way - the passage I had previously deleted. It reads as follows:

> Such a series of causes must however stop somewhere; for in it an earlier member causes an intermediate and the intermediate a last (whether the intermediate be one or many). Now if you eliminate a cause you also eliminate its effects, so that you cannot have a last cause, nor an intermediate one, unless you have a first. Given therefore no stop in the series of causes, and hence no first cause, there would be no intermediate causes either, and no last effect, and this would be an open mistake. (Ia, 2, 3)[17]

On my reading, Ia, 46, 2, 7 would limit this argument to efficient causation *per se*, and that, I think, is the right way to see the relationship between the core argument of the Second Way and the discussion of efficient causality in Ia, 46, 2, 7. The suggestion that Ia, 46, 2, 7 concerns a contrast between simultaneous and temporally prior efficient causality is textually unfounded, and thus its use to impose of a vertical interpretation on the Second Way wholly arbitrary.

If textual considerations do not favor a vertical reading of the Second Way over the horizontal reading, are there philosophical reasons involving the principle of charity that favor the vertical reading? I cannot see that there are. Read either way, the argument depends upon the claim that an infinite chain of dependence is repugnant to reason. For my own part, I cannot see why an infinite chain of efficient causes that stretches into the heavens simultaneously should be more repugnant to reason than an infinite chain of efficient causes stretching infinitely far back in time. Part of the difficulty in making this comparison is that it is hard to see what commentators have in mind when they speak of a hierarchy of simultaneous efficient causes. They are certainly not forthcoming with detailed illustrations. Perhaps the following example adapted from Kant will serve our purposes. A ball resting on a cushion is said to be the simultaneous cause of the hollow it makes in that cushion. Now let us replace the ball with another cushion that will make at least some depression in the cushion below it. On top of that cushion we place another, then another, and so on *ad infinitum*.

Of course, there are *physical* reasons why such an unending stack of pushed-downward-downward-pushing pillows is impossible, but there doesn't seem to be any reason to suppose that the bottom cushion could not be pushed down upon unless there exists a topmost pillow that is not itself pushed downward. Admittedly, there may be independent reasons for rejecting the possibility of an *actual* infinity of stacked pillows, but these may apply equally well against both the horizontal and vertical infinite regresses. As far as I can tell, then, there are no reasons derived from the principle of charity that should lead us to replace the supposedly naive hori-

zontal reading of the Second Way with the more sophisticated vertical reading.

Now my general thesis that the Five Ways present *vindications* rather than *demonstrations* of God's existence is, strictly speaking, independent of a horizontal rather than a vertical reading of the Second Way. Yet they are connected in the following important way. Read as an attempted demonstration of God's existence and interpreted horizontally, the Second Way may seem to demonstrate (or try to demonstrate) that at some time in the past the world was created by God. This would be embarrassing, for Aquinas is unequivocal in asserting "that the world has not always existed cannot be demonstratively proved but is held by faith alone" (Ia, 46, Reply). But this embarrassment disappears if we give up the idea that the Five Ways in general, and the Second Way in particular, are intended to demonstrate God's existence using premises derived from the natural sciences. Conversely, if we interpret the Second Way as a vindication of traditional religious belief in the face of a threat from (supposedly adequate) natural explanations in terms of efficient causes, then it surely concerns horizontal (i.e., temporally prior) causes, since this is where the threat comes from. For, of course, the view that the world has always existed and is wholly determined by temporally antecedent efficient causes is flatly *in*compatible with the Christian doctrine of creation. Thus Aquinas, who is generally committed to the compatibility of faith and reason, must reject all arguments intended to establish the eternality of anything save God.[18]

The First Way

Whereas the Second Way concerns efficient causes, the First concerns motion—in the very broad sense in which the Aristotelian tradition uses this concept. A key feature of this argument is the following principle:

> [To] cause change is to bring into being what was previously only about to be, and this can only be done by something that already is ...

That is, only something *actual* can cause the *potentially F* to become *actually F*. Aquinas illustrates this principle with the following example:

> Thus fire, which is actually hot, causes wood, which is able to be hot, to become actually hot, and in this way causes change in the wood (Ia, 2, 3).

This example has led a number of commentators to read the principle as saying, in Kenny's words, that "only what is actually *F* will make something else become *F*."[19] But this is not what Aquinas says. He simply says that for something to pass from being potentially *F* to being actually *F* this must be caused by something else that is actual. *He does not say that the cause must actually be F.* Furthermore, this cannot be what Aquinas means. God, the ultimate cause of all change is, according to Aquinas, unchangeable.[20] God, then, is not only an unmoved mover in the sense of a mover that is not itself moved, but also an unmoving mover.

Having gotten this straight, we are back on familiar ground. Those who hold that "natural effects are explained by natural causes" can only account for change by citing an endless ungrounded sequence of moved movers. To avoid this affront to our brute cosmological instincts (which some people have), we must transcend the natural order and acknowledge the existence of an unchanging being that is the first cause of change and is not itself changed by anything.

The Third Way

The Third Way concerns contingency and necessity. I do not fully understand what Aquinas means by these notions, but I am quite sure that we will misread him if we give them a contemporary (modal logician's) interpretation. The difference between Aquinas's use of these notions and a contemporary interpretation comes out in the following crucial passage:

> Some of the things we come across can be but need not be, for we find them springing up and dying away, thus sometimes in being and sometimes not. Now everything cannot be like this, for a thing that need not be, once was not; and if everything need not be, once upon a time there was nothing. (Ia, 2, 3)

On its face, this passage seems to involve a simple scope fallacy that goes something like this: If for every contingent being there is some time at which it does not exist, then there is some time at which no contingent being exists. Now I am sure that commentators have found ways of giving this argument a charitable reconstruction, but I do not want to talk

about that. For the present purposes, it is more important to notice that for Aquinas—unlike the contemporary modal logician—the contingent (the non-necessary) has temporal limitation built into its nature. Termination is an essential feature of contingent beings.[21]

With these thoughts in mind, we can return to those who think that all "natural effects can be explained by natural causes." If we assume that natural things possess only contingent existence and find some way of rendering plausible the argument that if everything were contingent then, at some time in the past, nothing would have existed, we can reach Aquinas's conclusion with (relatively) little difficulty:

> But if that were true [i.e., that once upon a time there was nothing] there would be nothing even now, because something that does not exist can only be brought into being by something already existing. ... Not everything therefore is the sort of thing that need not be; there has got to be something that must be. etc.[22]

Good or bad, it is at least clear what this argument is intended to show. It is not presented as an *a posteriori* demonstration of God's existence based upon natural principles. To the contrary, it is an effort to show that the very existence of a natural contingent world is inexplicable on the basis of natural principles alone.

CONCLUSION

The central thesis Kenny's *The Five Ways* is that "it is much more difficult than at first appears to separate [the Five Ways] from their background in medieval cosmology," and for that reason, "any contemporary cosmological argument would have to be much more different from the arguments of Aquinas than scholastic modernizations customarily are." (3-4) To simplify, Kenny argues—and his book is an attempt to establish the point in detail—that the science of Aquinas's day is so radically different in its underlying concepts from the science of our day that it is impossible to transpose those arguments into the idiom of modern science.

It should be clear that if my reading of the Five Ways is correct, then Kenny's criticisms are out of focus. On the present reading, medieval principles of cosmology do not appear as *premises* in proposed demonstrations of God's existence. Thus they do not stand in need of replacement. Furthermore, we can imagine those engaged in what Kenny calls "scholastic modernizations" of the Five Ways attempting to produce arguments against the pretensions of modern science precisely in the style that Aquinas used with respect to the science of his day. As Kant saw, the longings that produce cosmological reasoning have their source in the felt dissatisfaction with any natural explanation of the world. The science of our day is better science than that of Aquinas's, but to a person driven by the ideals of reason, with its demands for unconditioned completeness, both sciences will seem unsatisfactory—and for the very same reasons.

Dartmouth College

Received October 23, 1989

NOTES

1. Anthony Kenny, *The Five Ways* (London: Routledge & Kegan Paul, 1969).

2. See F. C. Copleston, *Aquinas* (Harmondsworth: Penguin Books, 1955), pp. 14-30.

3. See È. Gilson, *The Philosophy of St. Thomas Aquinas*, translated by Edward Bullough (Cambridge: W. Heffer & Sons, 1924), pp. 36-75.

4. Unless otherwise indicated, all citations are to the *Summa Theologica* (London: Blackfriars, 1964).

5. It is this reading that gives Kenny's basic criticism its point.

6. Following the tradition, we may, if we like, speak of these arguments as being *a posteriori* in at least two ways: they all depend upon observed facts and in a broad sense of "cause," they all involve inferences back from effects to causes. My claim is that they are not, however, *a posteriori* in the sense of being empirical arguments. They are not exercises in natural theology. At least that is what I will attempt to show.

7. See Kenny, pp. 71-95.

8. Fr. 1476 ^{b}22-24.

9. The Latin reads "*causæ efficientis.*" Perhaps the translator did not translate the word "*efficientis*" because it is clear from context that efficient causes are at issue.

10. How God functions as an efficient cause is another matter. We will come back to this.

11. A similar claim is made by Gilson, pp. 56-58.

12. The translation, including deletions, is Kenny's, pp. 41-42.

13. The bold printed passages were deleted by Kenny and restored from the Blackfriars translation. The remaining passages preserve Kenny's translation.

14. It should go without saying that if the man waving the political sign caused the disturbance, though not by waving a political sign, then it would still be *true*, though perhaps misleading, to say that he, referred to as the waver of a political sign, caused the riot. We can also imagine contexts in which wearing a grey shirt could provoke a riot (the hated Grey Shirts).

15. I'm not, of course, suggesting a deep similarity between Aquinas and Davidson. Aquinas was not, for example, an anomalous monist. On the other hand, some of the occasionalist Moslem theologians whom Aquinas attacks held views that seem Davidsonian in at least the following, admittedly remote, respect. Where Davidson denies the existence of psychological laws, these occasionalists denied the existence of natural laws or natural causes, reasoning that since God causes everything, "fire would not heat, but God would cause heat at the presence of fire." For Aquinas's response to this view, see, for example, *Summa Contra Gentiles,* III, 64-70. Aquinas is thus involved in a twofold task. On the one side, he is attempting to protect religious belief from the presumptions of natural science. That, if I am right, is the point of the so-called Five Ways. On the other side, he wishes to grant the natural sciences a domain of autonomy, provided, of course, that its practitioners do not pridefully reject religion in its favor.

16. Similarly, it is not in virtue of being one hammer in a series of hammers that a particular hammer is efficacious, thus an infinite regress here is, again, not an affront to reason.

17. On its face, this argument hardly seems persuasive, since a parallel line of reasoning would yield the result that there must be a first *left-hand* member of the series: ..., -3, -2, -1.

18. At Ia. 46, 1, Aquinas explicitly attacks arguments intended to show that creatures have always existed. Here he rather desperately tries to get Aristotle on his side even in the face of apparently explicit texts to the contrary, e.g., *Physics* VIII, 1 & 2.

19. Kenny, p. 21.

20. This is stated explicitly in Aquinas's *Reply to the Second Objection*, where he speaks of God as "an original cause which in unchangeable and necessary *per se*."

21. For most contingent beings there is both a time when they have not yet come into existence and a time at which they will no longer exist. The only exceptions are those contingent beings whose existence commenced at the moment of creation, since, according to Aquinas, time did not exist before creation. (See Ia. 46, 3.) So, strictly speaking, Aquinas is wrong in saying that anything "that need not be, once was not." Since, however, these exceptions presuppose the existence of God, admitting them does not weaken Aquinas's argument.

22. Citing the reasoning in the Second Way, Aquinas goes on to argue that an infinite regress of necessary beings is impossible.

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 27, Number 4, October 1990

DIRECTIONS OF JUSTIFICATION IN THE NEGATIVE-POSITIVE DUTY DEBATE

H. M. Malm

MY aim, ultimately, is to suggest a change in how we account for our evidence in favor of a morally significant difference between duties not to cause harm and duties to prevent harm. Our evidence is our considered moral judgments about particular cases—judgments indicating that acts of causing harm are worse than failures to prevent harm in at least some situations in which other things are equal. Accounting for our evidence requires that we explain *why* the acts differ in status. One common way to do this is to attribute the difference in the moral status of the acts to an in itself significant difference between the act-types. For example, we might claim that one act is worse than the other because acting is in itself worse than refraining, because violating a right is in itself worse than not violating a right, or more generally, because causing harm is in itself worse than failing to prevent harm.[1]

Accounts such as these share what I will call an *inward direction of justification.* They turn inward to the properties of the act-types when accounting for the difference in the moral status of the acts, as opposed to outward to the duties that proscribe the acts. (I will clarify this more in a moment.) The resulting structure of justification for an inward direction of justification is given in the following diagram.

Level 1: Causing harm and failing to prevent harm (the act-types).
↓
Level 2: Instances of causing harm and failing to prevent harm when other things are equal.
↓
Level 3: Duties not to cause harm and duties to prevent harm.

On this structure, the crucial difference lies at level 1. A significant difference at this level accounts for a difference in the status of the acts at level 2, and in turn a significant difference between the duties at level 3. Conversely, the lack of a significant difference at level 1 entails that there is no difference in the status of the acts at level 2, and no difference between the duties at level 3. A structure of justification, to be more precise, specifies the justificatory relationship between differences at the various levels, while a direction of justification specifies the way in which we move between claims at the various levels. For example, an inward direction of justification begins with our judgments at level 2, takes these first as evidence of an in itself significant different at level 1, and then moves back down the line to defend a principle at level 3.

Section I will discuss the role that the inward direction of justification (and its corresponding structure) plays in the debate about the moral significance of the difference between duties not to cause harm and duties to prevent harm. It will be seen that it gives rise to a number of problems. Section II will consider two ways to avoid these problems, and argue that the most promising way requires that we abandon the inward direction of justification and its corresponding structure.

I

Let us refer to duties not to cause harm as *negative duties*, and to duties to prevent harm as *positive duties.* Michael Gorr characterizes the negative-positive duty debate when he poses the following question:

> All else equal, is violating a negative duty not to cause harm
>
> a) morally worse than, or
>
> b) morally no different than, violating a positive duty to prevent an equivalent harm?

Those who answer that there is no moral difference between the two cases will be said to endorse the Moral Symmetry Principle, while those who consider violating a negative duty to be worse than violating its posi-

tive correlate will be said to accept the Moral Asymmetry Principle.[2]

This passage brings out two characteristics of the current debate. First, the standard method for determining whether there is a morally significant difference between the duties is to evaluate particular violations of the duties in circumstances in which other things are equal.[3] Judgments that the violations deserve different moral assessments provide evidence that the difference between the duties is significant. Judgments that the violations deserve equivalent moral assessments provide evidence that the difference is not significant. Second, there seems to be just two plausible principles to hold. One, the Moral Asymmetry Principle, entails that violations of negative duties are worse than violations of correlate positive duties when other things are equal. The other, the Moral Symmetry Principle, entails that violations of negative duties are equivalent to violations of correlate positive duties when other things are equal. Indeed, the very names "Moral Symmetry Principle" and "Moral Asymmetry Principle" suggest a contradiction.

The principles themselves, of course, are only contrary. In addition to the Moral Asymmetry Principle defined above, we might have asymmetry principles entailing any of the following (and more).

(i) violations of positive duties are worse than violations of correlate negative duties when other things are equal.

(ii) violations of positive duties are sometimes worse than, and sometimes equivalent to, violations of correlate negative duties when other things are equal.

(iii) violations of negative duties are sometimes worse than, and sometimes equivalent to, violations of correlate positive duties when other things are equal.

It is easy to explain why asymmetry principles entailing (i) and (ii) are not serious contenders for a principle about the moral significance of the difference between the duties. We simply have no evidence that violations of positive duties are ever worse than violations of correlate negative duties when other things are equal. But the same cannot be said about a principle entailing (iii). It would be supported by all the judgments that support the Moral Symmetry Principle *and* all the judgments that support the Moral Asymmetry Principle. Let us call this third principle the *Weak Asymmetry Principle*, for contrast, renaming the Moral Asymmetry Principle the *Strong Asymmetry Principle*.

The Weak Asymmetry Principle has generally been ignored. This is unfortunate for reasons to be discussed later. The reason it has been ignored, it seems, is that people have adopted an inward direction of justification when accounting for the evidence in favor of a moral difference between the duties. Using this direction of justification together with one further assumption, we can derive the conditions of the Strong Asymmetry Principle from the rejection of the Moral Symmetry Principle as follows.

The Moral Symmetry Principle entails that violations of correlate duties deserve equivalent moral assessments when other things are equal: the acts will be both wrong, both permissible, or both right, and the agents will be equally blameworthy, praiseworthy, or excused. Thus the minimally sufficient condition for a moral difference between the duties is one pair of cases in which other things are equal and violations of the duties deserve different moral assessments. Let us suppose that there is such a pair. Let us also grant that a difference in the moral status of two acts must be attributable to some difference in the properties of those acts. Since other things are equal in our evidentiary cases, the difference in the moral status of those acts cannot be attributed to differences in such things as motives, consequences and risks to the agents. It must thus be attributable to a difference(s) in the properties that the acts have in virtue of being instances of causing harm and failing to prevent harm. Thus there must be some difference between the properties of the act-types causing harm and failing to prevent harm that carries enough moral weight to account for the difference in the status of the acts. In short, there must be some difference that is morally significant *in itself*. But if there is such a difference, then it will give rise to a difference in the moral status of the acts in every pair of cases in which other things are equal. Thus if the Moral Symmetry Principle is false, then the Strong Asymmetry Principle is true.

The above line of reasoning might be challenged at a number of places. The assumptions on which I will focus are these.

(S1) If there is a pair of cases in which other things are equal and an act of causing harm is worse than a

failure to prevent harm, then the source[4] of the difference in the moral status of the acts must be an in itself significant difference between causing harm and failing to prevent harm.

(S2) If the difference between causing harm and failing to prevent harm is morally significant in itself (or if the difference between a more specific pair of properties of the act-types causing harm and failing to prevent harm is morally significant in itself),[5] then there will be a difference in the moral status of instances of causing harm and failing to prevent harm in every pair of cases in which other things are equal.

The first assumption commits us to an inward direction of justification. It requires that we ground the difference in the moral status of the acts on an in itself significant difference between the act-types. Its plausibility rests on the belief that if other things are equal in our evidentiary cases, there can be no other basis for the difference in status. The second assumption, if the first is correct, renders the Weak Asymmetry Principle untenable. Though it may seem to be the more problematic of the two, I will later argue that the first assumption is the one that can and should be rejected. For now, let us call the combination of the two assumptions *Assumption S*, and consider some of the problems that arise if Assumption S is correct.

The first problem concerns the set of our considered moral judgments about particular cases. By "the set of our judgments" I mean each of our sets, not the set of your judgments and my judgments combined. The problem is that the set of our judgments is more in accordance with the Weak Asymmetry Principle than either the Strong Asymmetry Principle or the Moral Symmetry Principle. Let me explain.

Both the Strong Asymmetry Principle and Moral Symmetry Principle have been defended with intuitively persuasive cases. But the cases offered in defense of the one principle tend to be different in character than the cases offered in defense of the other. On the one hand, the Strong Asymmetry Principle is supported by cases in which the agents' (or agent's) negative and positive duties *conflict* with another morally relevant consideration. Sometimes the conflict is between the duties themselves, as when an agent must choose between killing one innocent person and allowing another innocent person to die. Here it seems clear that the agent's negative duty not to kill overrides her positive duty to prevent a death. Other times the conflict is between a negative or positive duty on the one hand, and the agent's own welfare on the other. Richard Trammell offers a pair of cases in which one agent must spend $1,000 to avoid killing a stranger, and the other agent must spend $1,000 to prevent a stranger's death.[6] Trammell's cases are even more compelling if we imagine that the cost to the agent, were she to fulfill her duty, is not monetary but physical, e.g., the loss of a limb. It is widely recognized that a risk of serious harm to the agent can permit a failure to prevent an even greater harm to an innocent other, but not the causing of an even greater harm. Still other times the conflict is between a negative or positive duty on the one hand, and special duty on the other, e.g., the duty to look after one's children.[7]

On the other hand, the Moral Symmetry Principle is supported by cases in which the agents' negative and positive duties do not conflict with another morally relevant consideration. Judith Lichtenberg asks us to compare a case in which the agent is told that if he pushes a certain button a person in another room will die, with a case in which he is told that if he does not push a certain button a person in another room will die. According to Lichtenberg, "it seems incontrovertible that the agent's duty to refrain from pushing the button in the first case is equal to his duty to push it in the second."[8] If by this she means, at least in part, that it would be as wrong to violate the one duty as it would be to violate the other, then I think we would agree (assuming equivalent motives, equivalent long term consequences, and no risks to the agent). Of course we may not *want* to agree, because we think that that judgment will commit us to similar judgments in every other pair of parallel cases (e.g., ones involving a significant risk to the agent), but that is true only if assumption S is correct. Thus the first problem is that the set of our judgments supports the Weak Asymmetry Principle, but Assumption S renders that principle untenable.[9]

The second problem, which is related to the first, concerns the logic of the matter. In the absence of Assumption S, the truth conditions for a morally significant difference between negative and positive duties require a difference in the moral status of violations of the duties in only one pair of cases in which other things are equal. But in the presence of Assumption S, they require a difference in the moral

status of the violations in every pair of cases in which other things are equal. When we combine this point with the first problem we get the result that Assumption S commits us to much more than we logically *need* be committed to, and intuitively *want* to be committed to, by the assertion of a morally significant difference between negative and positive duties.

The third problem concerns our *access* to the basis for the judgments we take as evidence. Assumption S requires that the source of the difference in the status of the acts be an in itself significant difference between causing harm and failing to prevent harm. Determining the possible candidates for that difference is not an easy task, at least in part because, as Gorr puts it,

> ...the concept of causation has proven so intractable to analysis. If the difference between causing harm and failing to prevent it were analyzable in terms of the presence or absence of some reasonably large set of distinguishable features $f_1,...f_n$, then the defender of [the Strong Asymmetry Principle] would at least have some potentially fruitful research programs to pursue.... But if, as it appears likely, the difference is not further analyzable (at least on the basis of our everyday understanding of causation), then things look much bleaker.[10]

The fact that we do not have a set of necessary and sufficient conditions for causation does not show that we cannot point to *some* distinguishing difference between causing harm and failing to prevent harm to serve as the basis for the difference in status. But it does suggest that the task of finding one is difficult.[11] The problem is that regardless of which difference we choose, we have at best only inferential evidence that that difference is morally significant in itself. No one directly intuits, for example, that acting is in itself worse than refraining. Instead, we infer that it is worse, if we do, from our judgments indicating that *this* instance of acting is worse than *that* instance of refraining. But our inferences must be made from a set of judgments that are not consistent in that regard (re: problem one). Thus Assumption S commits us to holding that the *basis* for the difference in the moral status of the acts, and more importantly, the basis for a moral *principle* about the duties proscribing the acts, be something about which we have only (a) indirect intuitive knowledge, and (b) intuitive knowledge in which we cannot be wholeheartedly intuitively convinced.

II

We need to reject Assumption S. That assumption, recall, has two parts.

(S1) If there is a pair of cases in which other things are equal and an act of causing harm is worse than a failure to prevent harm, then the source of the difference in the moral status of the acts must be an in itself significant difference between causing harm and failing to prevent harm.

(S2) If the difference between causing harm and failing to prevent harm is morally significant in itself, then there will be a difference in the moral status of instances of causing harm and failing to prevent harm in every pair of cases in which other things are equal.

Since the first two problems of the previous section, and part of the third, stem from S2's implication that there will always be a difference in the status of the acts (other things equal), let us first consider the possibility of retaining S1 and rejecting S2. Whether that can be done depends, of course, on what we mean by "morally significant in itself"—a notion that has received far more use than analysis.

One option that captures the spirit of the phrase is this:

> The difference between *X* and *Y* is morally significant in itself if and only if the fact that act *A* is an instance of *X* and act *B* is an instance of *Y*, is alone sufficient to effect (account for) a difference in the moral status of *A* and *B*.[12]

After all, were the difference not alone sufficient, but sufficient, say, only in conjunction with some other morally relevant property, then there would be little point to the ascription of "in itself." And were it never sufficient, alone or otherwise, there would be little point to saying that it is morally significant at all.

Yet this definition will not allow us to escape the problems of the previous section. In order for the difference between *X* and *Y* to be alone sufficient to effect a difference in the moral status of *A* and *B*, it must be the case that either *X* is in itself worse than *Y*, or *Y* is in itself worse than *X*. But if either of these is the case, then there will be a difference in the

moral status of instances of *X* and *Y* in every pair of cases in which other things are equal.

A second option that seems more promising is this:

> The difference between *X* and *Y* is morally significant in itself if and only if the fact that act *A* is an instance of *X* and act *B* is an instance of *Y* can effect a difference in the moral status of *A* and *B*.

If this definition[13] is to be an improvement on the first, the difference between the act-types must be such that it *sometimes* effects a difference in the status of the acts when other things are equal, and sometimes not, other things equal. But how is that possible? Let me state the problem in the form of a *reductio*. Suppose that the difference between *X* and *Y* effects a difference in the moral status of acts *A* and *B* in a pair of cases in which other things are equal. Also suppose that there is another pair of acts, *A'* and *B'*, which differ with respect to *X* and *Y*, other things are equal, and *A'* and *B'* do not differ in status. Since *A* and *B* differ in status while *A'* and *B'* do not, there must be some difference between the two *pairs* of cases that accounts for the difference between those two pairs. That is, *A* and *B* must each have a morally relevant feature(s) that *A'* and *B'* each lack, that accounts for the fact that *A* and *B* differ in status while *A'* and *B'* do not. Let us call it *P*. Since *P* is needed to account for the difference in status between *A* and *B* (for without it they would be equal in status), the source of the difference in status is not the difference between *X* and *Y*, but at best the difference between *X&P* and *Y&P* (e.g., the difference between causing harm when the only other option is to incur a harm oneself, and failing to prevent harm when the only other option is to incur a harm to oneself). Further, since *P* is not a feature of the act-types causing harm and failing to prevent harm (or else it would be shared by *A'* and *B'*), the source of the difference in status in our evidentiary cases is not an in itself significant difference between those act-types. And that is inconsistent with S1.

It may be objected that I have made too much of S1's requirement that the source of the difference in status be an in itself significant difference between causing harm and failing to prevent harm. We ought to be able to attribute the difference in status, it seems, to a combination of properties, even if not all of the properties are true of either causing harm or failing to prevent harm.

Though our judgments make this view attractive, we cannot simply state that the difference between *this* pair of combinations gives rise to a difference in status, while the difference between *that* pair of combinations does not. Our task is to give a moral basis for our judgments, not redescribe them. What we can do is modify S1 as follows (and adjust the wording of S2 to cover any pair of act-types):

> (S1') If there is pair of cases in which other things are equal and an act of causing harm is worse than a failure to prevent harm, then the source of the difference in the moral status of the acts must be an in itself significant difference between some pair of act-types of which the acts in question are instances.

This assumption allows us to attribute the difference in status to a combination of properties, by allowing us to define the relevant act-types in a way that includes those properties. Thus we might distinguish between the act-types *XP* and *YP*, *XQ* and *YQ*, *XRS* and *YRS*, and claim that those differences are morally significant in themselves. And since a claim that the difference between *XP* and *YP* is morally significant in itself, has no implications for acts that differ with respect to *XT* and *YT*, we can give a basis for our evidentiary judgments while denying that acts of causing harm and failing to prevent harm will always differ in status, other things equal.

The above method seems to establish the tenability of the Weak Asymmetry Principle while using an inward direction of justification. Yet problems remain. First, our judgments indicate that acts of causing harm are worse than failures to prevent harm in many sorts of cases. Thus if we need to define a new pair of act-types for each sort of case, we will not have one, or even a few grounding claims, but myriad in-itself-significant claims. This greatly aggravates the access problem that was discussed in the previous section. Second, since our act-type descriptions have been tailored to fit our considered moral judgments (distinguishing, for example, between *XPUV* and *YPUV* on the one hand, and *XPUW* and *YPUW* on the other, and claiming that only the former distinction is morally significant in itself) our grounding claims are still little more than restatements, mildly generalized, of our considered moral judgments. *Why*, we should still want to know, is the

difference between some pairs of act-types morally significant in itself, while the difference between other pairs is not? Finally, if our grounding claims are defined in terms of a variety of act-types, it is difficult to determine what will become of our duties. They have traditionally been defined in terms of causing harm and failing to prevent it.

Let us consider one final definition:

> The difference between *X* and *Y* is morally significant in itself if and only if instances of *X* and *Y* differ in moral status in some pairs of cases in which other things are equal.

This definition drops the troublesome notion of effecting a difference. Yet if we want to retain S1 it is the least plausible of the three. Suppose we are asked "Why do that acts differ in status in those cases in which they do?" and answer "Because the difference between causing harm and failing to prevent harm is morally significant in itself." With the above definition we are merely asserting that the acts sometimes differ in status because the acts sometimes differ in status—an answer which obviously fails to provide a basis for the difference in status. Moreover, the above definition is consistent with holding that the act-types are morally different *because* the particular acts differ in status. But that option runs contrary to S1.

None of the definitions I have considered allow us to retain S1 and reject S2. That, of course, does not show that a suitable definition cannot be found, but it does, I hope, cast doubt on our chances for success. Let us turn our attention to S1. It is false if an inward direction of justification is not the only possible direction of justification when accounting for our evidence that negative and positive duties are morally different. Further, if S1 is false, then we are not committed to the problematic implications of S2.

Evidence can work in at least two ways. *E* can be evidence for *F* because *E* is (believed to be) able to cause or otherwise account for *F*. The discovery of carcinogenic chemicals in a city's previously clean water well is evidence that the city's cancer rate will rise. Similarly, the fact that cigarette smokers have higher rates of lung cancer than nonsmokers is evidence that nonagenarians are less likely to be smokers than nonsmokers. *E* can also be evidence for *F* because *F* is (believed to be) able to cause or otherwise account for *E*. The fact that all my jewelry is missing is evidence that I have been robbed. Similarly, the fact that cigarette smokers have higher rates of lung cancer than nonsmokers is evidence that cigarette smoke is carcinogenic. An inward direction of justification (and its corresponding structure of justification) assumes the first sort of evidentiary relationship between our judgments about the moral status of particular acts and the duties that proscribe the acts. Our judgments that the acts differ in status provide evidence that the duties are morally different, because a difference in the status of the acts could account for a difference between the duties. The duties would be different, that is, because causing a harm is (sometimes) a more serious wrong than failing to prevent a harm. An *outward direction of justification* takes advantage of the second sort of evidentiary relationship. Our judgments provide evidence that the duties are different because a difference between the duties could account for the difference in the status of the acts.

In more detail, an outward direction of justification begins, as does the inward, with our considered moral judgments about particular cases. But rather than taking these judgments first as evidence of an in itself significant difference between the act-types (a difference which grounds the difference in the status of the acts and in turn the difference between the duties), it takes them directly as evidence of a significant difference between the duties. Further, rather than assuming that the duties are morally difference because the acts (and act-types) are morally different, it assumes that the acts (and act-types) are morally different because the duties are morally different. The resulting structure of justification for an outward direction of justification thus far looks like this.

Level 1: Negative and Positive duties.
↓
Level 2: Instances of causing harm and failing to prevent harm when other things are equal.
↓
Level 3: Causing harm and failing to prevent harm (the act-types)

On this structure, the level 1 difference between the duties accounts for the difference in the status of the acts at level 2, and in turn the significance of the difference between the act-types at level 3. The act-type difference is significant, that is, *because* the difference between the duties is morally signifi-

cant—its significance is derivative. We might define it as follows:

> The difference between X and Y is morally significant if and only if the fact that a given act is an instance of X rather than an instance of Y is relevant to determining the moral status of the act.

If the duties are different, then whether an act is an instance of causing harm or an instance failing to prevent harm is relevant to determining the moral status of the act because it determines which sort of duty to consult. But since it is not *the reason why* the act has one status rather than another (that point is determined, thus far, by the difference between the duties), the difficult task of finding a difference between the act-types to which we can point and say "*that* is the difference that is morally significant in itself" is obviated.[14]

Given the above structure, the Weak Asymmetry Principle is tenable just in case we can explain the level 1 difference between the duties in a way that does not always entail a difference in status at level 2. Notice, first, that we should not explain the level 1 difference in terms of the moral status of the acts at level 2—asserting, for example, that the duties are different in that violations of one sort are morally worse than violations of the other sort, other things equal. That claim merely tells us something about the *implications* of the level 1 difference, and not something about the level 1 difference itself that can account for the difference in status at level 2. Instead, we can account for our judgments if we explain the level 1 difference in terms of the reasons the duties admit as grounds for justified or excused violations. Positive duties, our judgments indicate, permit or excuse violations on grounds and (or) in circumstances that correlate negative duties do not.[15] For example, in situations involving a significant risk to the agent, this risk provides a reason that can justify (permit) the agent's failure to prevent a greater harm, but not her causing of a greater harm.

More generally, two duties are morally different if, were we to make a complete statement of each duty which lists the grounds and (or) circumstances under which violations of that duty are permissible (e.g., "Do not kill except for reasons, or except in circumstances *Q*, *R*, or *S*"), the lists for those duties would not be identical. Some reasons may occur on both lists (e.g., killing or letting die in self-defense, in defense of third parties, in order to avoid killing two persons), while other reasons would occur on only one (e.g., letting die in order to avoid a serious harm to oneself, in order to avoid letting two die, in order to fulfill an important special duty). Thus the particular acts differ in status, when they do, because the circumstances of the acts provide a reason that counts as an adequate reason for violating only one of the two duties.[16] And when there is no moral reason in support of either violation, this sort of difference between the duties clearly allows that the acts are both wrong and the agents are equally blameworthy.[17]

The above account represents the first part of an outward direction of justification. The possibility of such an account entails that Assumption S is false. A complete account explains the difference between the duties in terms of the fundamental values of the moral system in which the duties operate. Jean Beer Blumenfeld provides the basics of this step when she writes (though apparently supporting the Strong Asymmetry Principle):

> [I]f liberty is a significant value, competitive with welfare, then the Maximal system [the Moral Symmetry Principle] is mistaken....because it is a radical denial of the autonomy of persons.... On an intermediate system [the Strong Asymmetry Principle] we have both negative and positive duties with the negative outweighing the positive. On the minimal system we have no positive duties at all.[18]

Thus Blumenfeld rejects the maximal system, in which negative and positive duties are morally equivalent, on the grounds that it "takes as overriding the single value of welfare and entirely ignores competing claims of freedom." And she rejects the minimal system, in which there are no positive duties at all, on the grounds that it "ignores the claims of welfare." The intermediate system, in which negative duties are more stringent, in some sense, than positive duties, is to be preferred on the grounds that it allows "both liberty and welfare a place as ultimate values."[19]

It is not necessary, however, to ground the duties in different values. A social contract theorist, for example, might argue that both sorts of duties are grounded in autonomy. Negative duties might be accepted on the grounds that, while they decrease the range of permissible actions, they increase the value of one's autonomy within the restricted range: the less effort one must spend to protect oneself, the

more effort one may spend on pursuing one's goals. Positive duties, which further restrict the range of permissible actions, may be accepted as ways to "hedge our bets." We might consent to prevent another's harm when we can do so without sacrificing our own important aims, on the hopes that if we find ourselves in dire straights there will be at least one other person who can prevent our harm without sacrificing her important aims. But if the positive duties were as stringent as the negative ones, then, I think it could be argued, we would have failed in our attempt to protect our autonomy. Not only would we be required (on most accounts) to come to another's aid whenever the harm that we could prevent is greater than the harm that we would incur, but another person could legitimately cause us harm anytime the harm that she could prevent is *equal to* (as well as greater than) the harm that we would incur. Her choice between killing one innocent person, and letting another die, for example, would be indifferent.

Warren Quinn also supports the intermediate system, at least indirectly, in a discussion about the correlative issue of negative and positive rights (rights against harmful interference and rights to aid or support):

> In giving him the authority to have primary say over what may be done to him [by including negative rights], morality recognizes his existence as an individual with ends of his own—an independent *being*. Since that is what he is, he deserves this recognition. Were morality to withhold it, were it to allow us to kill or injure him whenever that would be collectively best, it would picture him not as a being in his own right, but as a cell in the collective whole....
>
> None of this, of course, denies the legitimate force of positive rights. They too are essential to the status we want as persons who matter. But negative rights, for the reasons I have been giving, define the terms of moral possibility. Their precedence is essential to the moral fact of our lives, minds, and bodies really being ours.[20]

Each of the above accounts maintains that the maximization of human welfare is not *the* fundamental aim of a moral system. Other things, such as autonomy and the status of persons, matter as well.[21] And if they do matter, then our positive duties to prevent harm will not be as stringent as our negative duties not to cause harm. The final structure of justification for an outward direction of justification is this:

Level 1: Fundamental values
↓
Level 2: Negative and positive duties
↓
Level 3: Instances of causing harm and failing to prevent harm when other things are equal.
↓
Level 4: Causing harm and failing to prevent harm

In summary, an account which employs an outward direction of justification clearly can avoid the first two problems, and at least part of the third problem, of an account which employs an inward direction of justification. By attributing the difference in the status of the acts to a difference between the duties, and by explaining the latter difference in terms of the reasons the duties admit as grounds for justified or excused violations, we are not committed to more than logically need be committed to, or intuitively want to be committed to, by the assertion of a morally significant difference between the duties. And while some might object that we have only indirect intuitive knowledge about the importance of such values as autonomy and human welfare, it is, at the very least, intuitive knowledge in which we can be wholeheartedly intuitively convinced.

Loyala University of Chicago

Received September 28, 1989

NOTES

1. See, for example, Raziel Abelson, "To Do or Let Happen," *American Philosophical Quarterly*, vol. 19 (1982), pp. 219-27. Michael Wreen, "Breathing a Little Life into a Distinction," *Philosophical Studies*, vol. 46 (1984), pp. 395-402 and O. H. Green, "Killing and Letting Die," *American Philosophical Quarterly*, vol. 17 (1980), pp. 195-204.

2. Michael Gorr, "Some Reflections on the Difference Between Negative and Positive Duties," *Tulane Studies in Philosophy: Positive and Negative Duties*, ed. by Eric Mack (New Orleans: Tulane University Press, 1986), p. 93.

3. The "other things equal" clause requires equivalences in all morally relevant considerations other than the possibly relevant distinction or consideration in question (e.g., equivalences in such things as motives, consequences and risks to the agent). However, in order to use our judgments about particular cases as evidence for or against a given distinction, and in order to give a plausible reading to a claim that a given distinction is significant, we need to restrict our attention further to *fair test cases*. Fair test cases are cases in which the implications of a given distinction, if it is significant, are not cancelled, or hidden from our view, by a factor that is present in both situations. For example, cases in which both agents are insane are not fair test cases when arguing that a given distinction has no relevance for the status of agents. Nor is it fair to interpret a claim entailing that one sort of conduct is morally more blameworthy than another sort of conduct, as requiring a difference in the blameworthiness when both agents are insane. Also to be avoided are cases in which a given feature, such as a gruesomely heinous motive, overwhelms our powers of judgment. (I discuss another possible condition of the "fair test case" restriction in note 5.) For brevity I will assume that the "fair test case" restriction is included in the "other things equal" clause.

4. "Source" and "effect" (which I will use later on) may be understood loosely, along the lines of "the basis for" or "the reason for" on the one hand, and "gives rise to" or even "justifies" on the other. I use them to help emphasize the relationship between claims at the various levels.

5. For brevity, I will concentrate on the difference between causing harm and failing to prevent harm, with the understanding that the significant difference may be this general one or a more specific difference in properties. However, if the significant difference is a more specific difference in properties, and *if* one of those properties is (as Francis Kamm says) "exportable" to acts of the other kind, then we need to add another condition to the "other things equal-fair test case" restriction. Suppose that X is one of the definitional properties of causing harm, but not one of the definitional properties of failing to prevent harm. Also suppose that X has intrinsic moral significance. If X is exportable to instances failing to prevent harm, that is, if it can be present in instances of failing to prevent harm, then in pairs of cases in which both acts have X (one by definition the other by circumstance), we cannot expect to see a difference in status. The "other things equal-fair test case" requirement will henceforth be assumed to exclude cases in which an intrinsically significant exportable property is exported. See Francis Myrna Kamm, "Harming, Not Aiding, and Positive Rights," *Philosophy & Public Affairs*, vol. 15, no. 1, 1986, for a detailed and insightful discussion about how we can equalize situations for and against exportable properties, and how we can use cases which are equalized for exportable properties (that is, theexportable properties are exported) to test the moral significance of those properties. It is worth noting, however, that if *X* is a definitional property of causing harm, and if *X* is exportable to instances of failing to prevent harm, then *-X* is *not* a definitional property of failing to prevent harm. Thus even if the difference between *X* and *-X* is morally significant in itself, it may be misleading to say that this shows or entails that the difference between causing harm and failing to prevent harm is morally significant in itself.

6. Richard Trammell, "Saving Life and Taking Life," *The Journal of Philosophy*, vol. 72 (1975), p. 131.

7. See, for example, Raziel Abelson, "To Do or Let Happen," *American Philosophical Quarterly*, vol. 19 (1982), p. 227.

8. Judith Lichtenberg, "The Moral Equivalence of Actions and Omissions," *Canadian Journal of Philosophy*, Supplementary Volume VIII (Toronto, 1982), p. 25.

9. Interestingly, there rarely seems to be disagreement about the proper assessment for a given case. Instead, disagreements tend to focus on the implications of the proper assessment for a given principle. For example, when faced with an assessment that seems inconsistent with a favored principle, we might argue that "other things" are not truly equal.

10. Gorr, "Some Reflections," *op. cit.*, p. 97. Gorr's focus at this point is on the possibility of finding a difference that one can convince one's critics is morally significant. My focus is on the difficulty of finding a difference that one can convince oneself is morally significant.

11. Some of the more widely discussed differences, such as those between acting and refraining, and violating a right and not violating a right (see note 1 for references), may not properly correlate with causing harm and failing to prevent it. In "Killing, Letting Die, and Simple Conflicts" (*Philosophy and Public Affairs*, vol. 18 (1989), I argue against the acting-refraining distinction, and in "Justice and Charity," *Ethics*, vol. 97 (1987) Allan Buchanan raises the possibility that persons have rights to have their harms prevented, rights that do not arise from a special relationship between the agent and potential victim.

12. Being *alone sufficient* does not entail that there *must* be a difference in the status of the acts, for there may be other differences between the two acts that counteract the effects of this one. For example, one act may have much more devastating consequences than the other.

13. These definitions are not definitions, strictly speaking. The first could be stated as follows: The difference between *X* and *Y* is morally significant in itself$=_{dfn}$ *X* is in itself worse than *Y* or *Y* is in itself worse than *X*. It is difficult to state the second and (upcoming) third options in the form of a definition, at least if we want the definition to (a) tell us something about *X* and *Y*, as opposed to the implications of *X* and *Y*, and (b) allow us to cite the difference between *X* and *Y* as the basis for the difference in status of the acts.

14. For example, attempts to adjust the senses of "acting" and "refraining" so that they correlate with causing harm and failing to prevent harm are pointless, at least if the aim of the adjustment is to allow us to say that the particular acts differ in status (or say that the act-types are morally different), *because* the difference between acting and refraining is morally significant in itself. And if the aim is to make a general claim that maps (but does not ground) our judgments, then we would do just as well to stick with "causing harm" and "failing to prevent harm."

Also, it might be objected that the difference between "violates a negative duty" and "violates a positive duty" could serve as the difference between the act-types that is morally significant in itself. Yet even if we cite this difference, the difference between the act-types would still *derive* its significance from the difference between the duties. Thus we cannot return to a structure in which the act-type difference plays a more fundamental role than the duty-difference. Further, if we are to defend the Weak Asymmetry Principle, we cannot claim that the bare difference between these properties is what accounts for the difference in the status of the acts.

15. I develop this view in detail in "Between the Horns of the Negative-Positive Duty Debate" (forthcoming in *Philosophical Studies*), arguing that negative duties are stricter than positive duties, when "stricter" is defined as follows:

Duty *N* is stricter than duty *P* if and only if, other things equal, (a) there are some reasons that can justify (excuse) a violation of *P* that cannot also justify (excuse to the same degree) a violation of *N*, and (b) there are no reasons that can justify (excuse) a violation of *N* that cannot also justify (excuse to the same degree) a violation of *P*, when (c) the reasons are relevant to both violations and the violations occur in relevantly similar situations.

That paper does not make clear, however, the structure of justification that this definition presupposes, nor does it establish the need for adopting this structure if we are to defend the Weak Asymmetry Principle in a plausible way.

16. This is a simplified account. It needs to be expanded to cover the status of the agents as well as the acts, as well as allow that even if a reason is not sufficient to permit or excuse either violation, it can unequally affect the status of those violations (e.g., by excusing one agent to a greater degree than the other).

17. In other words, the present account allows that failures to prevent a harm are *prima facie as bad as* acts of causing harm, but denies that being "*prima facie* as bad as" is equivalent to being morally indistinguishable. The acts are violations of morally different duties, and the difference between the duties only sometimes entails a difference in the status of the acts, other things equal.

18. Jean Beer Blumenfeld, "Causing Harm and Bringing Aid," *American Philosophical Quarterly*, vol. 18 (1981), p. 329. Blumenfeld seems to be viewing the intermediate system as entailing that violations of negative duties are worse than violations of correlate positive duties when other things are equal. She argues, for example, that our evidentiary judgments show that "killing *is worse* than letting die," and speaks of negative duties as being *stronger* than positive duties (emphasis added, p. 328). Regardless of whether she wants to defend the Strong Asymmetry Principle or the Weak Asymmetry Principle (both of which function on an intermediate system),her discussion about the values helps to explain why positive duties may be violated on grounds and (or) in circumstances in which negative duties may not.

19. Blumenfeld, *op. cit.*, p. 329.

20. Warren Quinn, "Actions, Intentions, and Consequences," *The Philosophical Review*, vol. 98 (1989), pp. 309-10.

21. A consequentialist might object that she too can maintain that autonomy matters. Granting persons rights over their own lives (and thus obliging others not to interfere) may be one way to help to maximize the general welfare. But, as Quinn argues, such a view is unlikely to "give us the moral image of ourselves we think fitting. For it locates the ultimate ground of proper deference to a person's will in the fact that such deference maximizes the general balance of good. In such a system, it is not so much his right to have his way that really matters as the general goodness of letting him have his way" (Quinn, *op. cit.*, p. 311).

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 27, Number 4, October 1990

NOMINALISM AND ABSTRACT REFERENCE

J. P. Moreland

MOST people would grant that certain sentences are true which appear to involve reference to universals. Among those sentences are these:

(1) Red resembles orange more than it resembles blue.

(2) Red is a color

(3) Humanity is a substance-kind.

The Realist has a relatively straightforward way of accounting for the truth of these sentences.[1] She can argue that the subject terms refer to a universal. This can be made explicit by the following paraphrases:

(1a) Redness resembles orangeness more than it resembles blueness.

(2a) Redness is a color.

(3a) Mankind is a substance-kind.

The Realist argues that there are states of affairs that obtain in the world which are accurately described by (1a)-(3a). This point can be made more linguistically by claiming that these sentences incorporate terms which refer to universals and, further, that these terms play essential roles in these sentences and cannot be eliminated through paraphrase. Thus, the truth of these sentences presupposes the existence of universals.[2] The terms that allegedly refer to universals in sentences (1a)-(3a) are called abstract singular terms, e.g., "redness," "orangeness," "blueness," and "mankind."

The Realist, then, challenges those who deny the existence of universals to account for sentences (1)-(3) in such a way that they still make plausible claims about the world without entailing the existence of universals. Now it is generally agreed that there are three major schools of thought regarding the nature of properties—Extreme Nominalism (properties do not exist), Nominalism (properties exist and are themselves particulars), and Realism (properties exist and are universals). Further, it is generally agreed that, whereas sentences which exhibit abstract reference like (1)-(3) are problematic for Extreme Nominalism, they can be adequately handled by Nominalism and, thus, abstract reference cannot be used as an argument for or against Nominalism *vis a vis* Realism.

The purpose of this article is to refute that claim. Specifically, I will show that sentences like (1)-(3) are equally troublesome for Nominalism and Extreme Nominalism, and that the phenomenon of abstract reference does lend support to a Realist interpretation of properties. First, a brief sketch of Extreme Nominalism, Nominalism, and Realism will be given. Second, we will investigate Extreme Nominalist and Nominalist arguments which seek to show that abstract singular terms refer to sets of concrete or abstract particulars. Finally, we will consider Extreme Nominalist and Nominalist attempts to offer reductive paraphrases of sentences which incorporate abstract singular terms.

I. EXTREME NOMINALISM, NOMINALISM, AND REALISM

A helpful way to compare and contrast different view about universals is to consider a case of quality-agreement. Let us limit our discussion to monadic, first-order properties. Suppose we have before us two round, red spots where each has the same infimae species of color and shape. Let us call our spots Socrates and Plato. How are we to account for this example of quality-agreement?

Extreme Nominalism is one answer to this question. Extreme Nominalists deny the existence of qualities altogether by giving them a reductive analysis as follows: a has the quality F$\leftrightarrow$P.

For example, *P* could be replaced by "*a* is a member of the set of *F*-things." Thus, Extreme Nominalists deny an ontology of qualities and quality-instances and only allow for concrete particulars—red spots, individual humans—and sets (predicates, concepts) of concrete particulars. The

color agreement between Socrates and Plato is to be explained by the fact that they are both members of the set of red-things. W. V. O. Quine and Wilfrid Sellars are examples of Extreme Nominalists.

A second school of thought is Nominalism. A Nominalist acknowledges the existence of qualities but denies that quality agreement is to be explained along Realist lines wherein qualities are taken to be universals. The Nominalist denies that the-redness-of-Socrates (red_1) and the-redness-of-Plato (red_2) is a numerically identical entity (the universal, redness) in each. Rather, each spot has its own little redness in it, *viz.* a particularized quality. Particularized qualities have been given a number of different names, among which are these: "tropes," "perfect particulars," "abstract particulars," "cases," "unit properties," "moments," and "quality instances." Advocates of Nominalism include D. C. Williams, G. F. Stout, and Keith Campbell.[3] In my view, Campbell is the most articulate contemporary proponent of Nominalism. He calls a particularized quality a grope, and since Campbell has worked out a Nominalist ontology more than anyone else in the current literature, it will be helpful later if we spell out his doctrine of tropes in more detail.[4]

According to Campbell, "Socrates is red" should be analyzed in this manner: The simple trope, red_1, which is a member of the set, redness—the set of all and only red tropes which stand to red_1 in the primitive relation of exact similarity—is a part of the compresent bundle of tropes, Socrates.

Three features of tropes bear special mention. First, a trope is a particular, that is, it is exhausted in one embodiment. It is not multiply exemplifiable. A trope is an infimae species taken as a particular. Second, a trope is a basic, primitive, simple entity. For Campbell, this means the following: a trope is simple in that it is not a whole with more basic parts in it, a trope is fundamental in that its existence is not dependent on anything else (e.g. Socrates' existence depends on the existence of red_1 but the existence of red_1 is not dependent on anything else), and a trope is independent in that a given trope could be the only thing in the universe. Tropes, then, are basic, simple entities which altogether lack complexity.

A third important feature of a trope is the central role that space (or space-time) plays in spelling out what a trope is. Tropes are essentially regional, that is, they exist at specific times and specific places. A trope is a quality-at-a-place. This should not be taken to imply, for example, that the redness of red_1 and the location of red_1 are two different constituents in red_1. Tropes are simple entities and not complex entities. Rather the place or formed-volume (Campbell uses these synonymously) of a trope like red_1 is identical to the redness of red_1; they differ only by a distinction of reason and not by a real distinction. The location and color of red_1 are merely different ways of thinking about the same trope and differ from one another only in thought.

The third view of qualities is Realism. Realists may differ over a number of issues, e.g. the existence of uninstantiated universals, but they are agreed on their analysis of quality-agreement. Socrates and Plato agree in color because each has a multiply exemplifiable entity, redness, which is literally in red_1 and red_2 as a constituent. Red_1 and red_2 are complex entities which have the same nature (redness) in each. Gustav Bergmann, Edwin B. Allaire, and D. M. Armstrong are proponents of Realism.[5]

It has already been pointed out that most philosophers who work on problems involving universals believe that Nominalism is superior to Extreme Nominalism when it comes to the phenomenon of abstract reference. D. C. Williams confidently asserted that "all the paradoxes which attend the fashionable effort to equate the universal Humanity, for example, with the class of concrete men...disappear when we equate it with our new set, the class of abstract humanities—the class whose members are not Socrates, Napoleon, and so forth, but the human trope in Socrates, the one in Napoleon, and so forth."[6] Nicholas Wolterstorff claims that "in general, it would seem that for every sentence containing singular terms standing for predicables, one can easily produce, as replacement, a sentence containing general terms true of cases [tropes] of predicables. All reference to predicables can be eliminated."[7] D. M. Armstrong and Michael Loux have made similar statements.[8] Let us see why these statements cannot be sustained.

II. Abstract Singular Terms Refer to Sets of Tropes

Let us recall sentence (1): (1) Red resembles orange more than it resembles blue.

Extreme Nominalism

How might an Extreme Nominalist explain the

truth of (1)?[9] An Extreme Nominalist could propose the following:

(1b) Anything red resembles anything orange more than it resembles anything blue.

But (1b) is not equivalent to (1). It could be the case that some red things resemble some blue things more than they resemble orange things. A red rubber ball would resemble a blue rubber ball more than it would resemble an orange tent. The problem for the Extreme Nominalist is that red things are complex in a way in which redness is not.[10] So there are other aspects in which red things could resemble blue things (size, shape, hardness, etc.) besides color.

An Extreme Nominalist could respond by asserting:

(1c) Any thing red color-resembles anything orange more than it color-resembles anything blue.

A number of problems can be raised against (1c). First, Realists like Armstrong have pointed out that "color-resembles" is likely to be a fabricated predicate like "believes-that-the-cat-is-on-the-mat," and not a predicate that is really primitive.[11] The predicate "color-resembles" does not stand on its own. A whole variety of shapes, smells, and other qualities will resemble one another as do red, orange, and blue. But if (1c) is correct, each of these cases of resemblance will require a new primitive predicate. But it seems clear that we understand the pattern that is common to all these predicates and this understanding is what permits us to form new resemblance-predicates in new cases. And this ability is best explained by arguing that what is common to these predicates is resemblance and what is unique is the respect of resemblance. Thus, the best way to understand the predicate "color-resembles" is to hold that when it is true that x color-resembles y, then this means that x resembles y in color.

A second objection to (1c) can be brought out by considering a possible world where "red" and "triangular" are co-extensive, "orange" and "sweet" are co-extensive, and "blue" and "square" are co-extensive. In this world, anything triangular color-resembles anything sweet more than it does anything square. But it would not be the case that triangularity would resemble sweetness more than it does squareness. There is more to red resembling orange more than it does blue than what (1c) captures, and thus, (1c) is inadequate.

Trope Nominalism

Armstrong argues that these arguments succeed against Extreme Nominalism (he calls it Orthodox Nominalism), but not against Nominalism (he calls it Particularism): "We see, then, that the arguments of this section, though very powerful against orthodox Nominalism, fail against Nominalism when it is combined with Particularism."[22] But this is not the case.

Let us first consider Nominalist paraphrases that take "red," "orange," and "blue" to refer to sets, that is, sets of tropes. The Nominalist might offer the following:

(1d) The set composed of red tropes resembles the set composed of orange tropes more than it resembles the set of blue tropes.

By using (1d), the Nominalist treats "red," "orange," and "blue" as singular referring terms which name three different sets of tropes.[13] But is it true to say that these three sets (not their members, but the sets themselves) resemble one another in the way that red, orange, and blue do in (1)? If sets exist at all, they are not colors or qualities. Two sets may resemble one another in being sets, in being abstract objects, in having members, and the like. But these respects of resemblance have nothing to do with color. Sets are just not the sorts of things that are colored.

In this regard, consider the following case. Suppose that the number of red and blue tropes were the same and the number of orange tropes were half this number. Then the set of red tropes would resemble the set of blue tropes (in the respect of having n members) more than it would resemble the set of orange tropes. But redness would not resemble blueness more than orangeness. So (1) and (1d) are not the same.

The Nominalist cannot respond by using the primitive predicate "color-resembles." Sets do not "color-resemble" one another. And there would still be the problem raised against Extreme Nominalism regarding the artificial nature of this predicate. The Nominalist could argue that (1d) is using terms which refer not to the sets themselves, but to the various members of those sets. But this move shifts strategies. Now "red," "orange," and "blue" are not abstract singular terms, but rather general terms which serve as abbreviations for discourse about

various tropes. We will consider this strategy in the next section.

Perhaps enough has been said about (1d). Let us move on to another Nominalist paraphrase which attempts to avoid the objections that have been raised against (1d). Consider this paraphrase:

(1e) The set composed of red tropes and the set composed of orange tropes are co-members of more natural sets than are the set composed of red tropes and the set composed of blue tropes.

This paraphrase is an attempt to remove an appeal to resemblance and substitute for it some reference to set membership. Sentence (1d) failed, in part, because it substituted sets for Realist qualities in (1), but kept resemblance intact. So (1e) is an attempt to be consistent. It substitutes sets for Realist qualities and set membership for resemblance.

This paraphrase also tries to capture a certain group of facts that the Realist captures in (1a). For example, the set of red tropes might itself be a member of some higher order set, say, the set of sets whose elements occur in the lower end of the color spectrum. Since orange is, but blue is not in this end of the spectrum, then the set of orange tropes would also be a member of this set and the set of blue tropes would not. The word "natural" in (1e) is an attempt to capture the fact that certain elements, as an ultimate, brute fact, simply fall into certain sets and others do not. If "natural" was not added, then (1e) would clearly fail. For there could be an indefinite number of arbitrary sets in which the set of red tropes and the set of blue tropes were members. Likewise, with the set of red tropes and the set of orange tropes. In that case, there would be no way of saying the set of red tropes and the set of orange tropes were co-members of more sets than the set of blue tropes.

Unfortunately, (1e) fails as an adequate paraphrase.[14] First, a set's identity is in its members. A set cannot change membership and be the same set since a necessary feature of a set is that it have just the members it does, in fact, have. However, consider the possibility that there might have been more red tropes than there tenselessly are. In that case, the new set of red tropes would be different from the one referred to in (1e). Sentence (1e) would use the term "the set composed of red tropes" to refer to a different entity than would be the case if (1e) were used to refer to the set of red tropes that tenselessly obtain in the actual world. But, then, (1e) would be describing a different state of affairs than originally designated. But nothing would have changed regarding the relations of resemblance between redness, orangeness, and blueness for these resemblance relations do not vary with the number of instances of redness, orangeness, or blueness.

In order to avoid this problem it would seem that the Nominalist would have to hold that it is an essential feature of redness that it have just the members it does. But, then, since red objects are red in virtue of having red tropes in them, this would amount to the claim that there could not have been more or less red objects. But this is absurd.

Second, consider a world where all the red tropes are co-present with strawberry taste tropes, all blue tropes are co-present with raspberry taste tropes, and all orange tropes are co-present with the taste tropes of a bitter orange flower. Recall that a trope is a simple entity wherein the nature and location of a trope are identical—they differ only by a distinction of reason. By the transitivity of identity, if the color of a trope is identical to its location, and if the taste of a trope is identical to its location, then if the two tropes are at the same location, they must be identical.

In the world just described, the red tropes would be identical to the strawberry taste tropes and, likewise, the blue and orange tropes would be identical to the raspberry taste tropes and the bitter orange tropes, respectively. But, then, the set of red tropes and the set of blue tropes could be co-members of more natural sets (say, the set of sets whose members are tart tasting tropes) than either would be with the set of orange tropes (which would be in the set of sets whose members are bitter tasting tropes). But redness would not resemble blueness more than orangeness.

This same objection could be raised in a world where red tropes are located nearer blue tropes than orange tropes. In this case, the set of red tropes and the set of blue tropes could be co-members of more natural sets (say, the set of sets whose members were in a certain spatial region) than the set of orange tropes.

Third, suppose there were a world where there was a missing shade of color between red and orange, e.g. red-orange or pink. In this world there would be no red-orange or pink tropes. In this case, the set of sets whose members were at the lower end

of the spectrum would be a different set, since it would not have as members the set of red-orange tropes or the set of pink tropes. Now, it may still be the case that the set of red tropes would be co-members with the set of orange tropes in more natural sets than would be the case between either and the set of blue tropes, although examples could be set up where this would not be true (by postulating a world where a number of color tropes were absent).

Nevertheless, (1e) spells out resemblance as co-membership in natural sets, and in this possible world, there would be less natural sets in which the set of red tropes and the set of orange tropes would be co-members. Thus, red and orange would bear a different resemblance to one another in this world than they do in the actual world because resemblance is treated by the Nominalist as co-membership in sets. But how can the presence or absence of red-orange or pink have any bearing on the resemblance between red and orange?[15]

In summary, it seems that the Nominalist strategy of paraphrasing (1) in terms of abstract singular terms referring to sets fails. There may be other paraphrases a Nominalist could offer. But I believe these would fail like (1d) and (1e) for three reasons. First, universal qualities are just not sets, and if sets do, in fact, exist and resemble one another, then the way sets resemble is not the way red, orange, and blue resemble. Sets are not colors. Second, universals such as redness, orangeness, and blueness and their resemblances are features of the world that obtain independently of the number of instances of those or other universals (such as red-orange or pink). It is not so with sets. Sets do depend on the number of their members. Third, the resemblance among universals expressed in (1) are resemblances in one respect, *viz.*, color. But tropes resemble in more than one respect (each has a "nature" and a "location" in some sense, and can resemble other tropes in either way). Thus, they can be grouped into resemblance sets in more than one way.

Perhaps a Nominalist could try a second approach, namely, a reductive approach. Here the Nominalist could argue that "red," "orange," and "blue" in (1) are not singular terms, but rather, abbreviations for discourse about a plurality of tropes. I will discuss this strategy as it applies to sentences like (2) Red is a color.

This is merely for convenience, because most discussions of the reductive approach in the literature focus on sentences like (2).

III. Abstract Singular Terms and Reductive Paraphrases

Extreme Nominalism

How might an Extreme Nominalist paraphrase (2). He might try the following: (2b) Everything red is colored.

This will not work. Consider the scattered location, *L*, of all red things. Everything *L*-located is colored, but *L*-location is not a color. Similarly, everything red might have been triangular so that everything triangular was colored. But triangularity would not be a color. The Extreme Nominalist could try, instead, (2c) Necessarily, everything red is colored.

This could avoid the objections raised against (2b). But it is still inadequate, for everything red is also extended, shaped, and located and, arguably, this is true necessarily. But redness is not extension, shape, or location.

Nominalism

Philosophers like Wolterstorff, Armstrong, and Loux believe that Nominalist's can adequately handle sentences like (2) by employing a reductive strategy. Again, I disagree. The Nominalist paraphrase that is usually offered as being a successful treatment of (2) is something like this: (2d) Reds are colors.

Sometimes (2d) is expressed by saying, "Everything which is a red is a color." Armstrong paraphrases (2) like this: "For all (ordinary) particulars, *x*, if *x* has a (particular) colouredness and the class of particular rednesses is a sub-class of the class of the particular colouredness."[16] There are three categories of arguments I want to raise against (2d).

1. *Arguments Against Extreme Nominalism Applied to Nominalism*

Recall the objection raised against the Extreme Nominalist paraphrase (2b). There it was pointed out that the scattered location, *L*, of all red things is not itself a color even though everything *L*-located is a color. Now, this same objection applies to (2d). Tropes are necessarily located at the place where they exist (and they are extended and shaped as

well). Consider the scattered location, L', of all red tropes. Everything L'-located would be a color, but L'-location is not a color. Since the location (which Campbell identifies with formed-volume) and nature of a trope are identical—they differ only by a distinction of reason—then we can put this as follows: L'-locations are colors.

One could object that there may be other tropes at some of the same locations as some red tropes and, therefore, it is false that everything L'-located is a color. But two things can be said against this idea. First, even if it were true, this would merely be a contingent fact. Tropes are independent entities—the very alphabet of being is what D. C. Williams called them—and all red tropes could have existed at locations where no other tropes were present. Second, since all tropes have natures which differ from their locations by a distinction of reason, then their natures are identical to their location. It follows that all tropes at the same location are actually identical. In fact, tropes can arguably be treated as bare particulars identical to places. It would seem, then, that the Nominalist doctrine of the simplicity of tropes leads to inconsistencies when it comes to permitting the compresence of tropes.

What about a Nominalist paraphrase of (2d)? A Nominalist might try (2e) Necessarily, reds are colors.

The problem is that it is also necessary that reds are extended (as well as shaped and located), so one could say (2f) Necessarily, reds are extensions.

But the universal, redness, is not an extension (or a shape or location). In response, a Nominalist might appeal to Armstrong's version of (2f)[17]

> (2f') For all (ordinary) particulars, x, if x has a (particular) redness, then x has a (particular) extendedness and it is not the case that the class of the particular rednesses is a sub-class of the class of the particular extendednesses.

But this is not true. The red tropes are necessarily at their location, i.e., have their particular extendedness (extension and location are identical). Indeed, red tropes are identical to their locations. So the class of particular rednesses is identical to the class of their particular extendednesses (they have the same members), and, therefore, the class of particular rednesses is a sub-class of the class of particular extendednesses. So it seems that some of the objections raised against Extreme Nominalism are equally effective against Nominalism.

2. *Arguments Based On a Realist Account of Higher Order Predication*

There is a further problem which can be raised against (2d). This can be brought out by thinking about what it means to say: (2) Red is a color. Most Realists are agreed that (2) is an example of predication, specifically, the predication of a second order universal of a first order universal. But Realists have differed as to how to analyze (2).[18]

For the moment, let us assume that the following account of (2) is the best one among Realist alternatives. In this account, (2) expresses a genus/species relation in the category of quality. This relation is one of essential predication. In other words, (2) says that color (a second order universal), an identical constituent of all first order colors, is an essential constituent of redness. If one affirmed that blue is a color, one would be saying that color is a constituent of blueness. Thus, in cases of inexact resemblance, say when red resembles blue, you have a generic identity (the second order universal, color, is a literal, essential constituent in redness and blueness) and a specific diversity (red is this-color, blue is that-color).

If this account is correct, then (2) would be better expressed as (2') Red is color. (2') asserts that color is an essential constituent in red and it is predicated of some other constituent in red (a bare particular, perhaps), and red is identical to this-color. (2') becomes the ontological ground for (2). (2) contains an "is" of classification.[19] As it stands, (2) says that red is a color. It places red in the class of colors and (2') grounds this classification by saying that red has a constituent in it, color, which is that entity which all members of the class of colors share in common.[20]

From what has just been said, it follows that a proper account of (2) involves three things: 1) seeing that (2) expresses (or presupposes) essential predication, 2) grounding the membership of red in the class of colors, and 3) assaying redness as a complex entity, *viz.*, this-color, where color is the genus (nature, essence) of redness. If this is what a proper account of (2) involves, then neither (2d) nor any other Nominalist paraphrase of (2) will succeed, for they all must treat red as a simple entity.

This point was brought out long ago by J. R. Jones against the Nominalists of his day like G. F. Stout.[21]

Jones argued that if x is a trope, then since x is simple, any attempt to assay the constituents of x that involves a predication such as Fx & Gx cannot be true. Fx & Gx says that F and G are constituents of x. Now let x be a red trope, say red_1. It would seem that (2d) says of each red trope that it has color as a constituent. In other words, if F is red and G is color, then (2d) appears to say that red_1 has color as a constituent. Red_1 has redness and it has color. So red_1 is not simple.

To avoid this conclusion, the Nominalist must hold that (2d) does not state that a trope has its redness or its color. (2d) becomes a way of stating the brute, unanalyzable fact that red tropes are in a set that is itself in the set of colors. But this is not a paraphrase of a Realist account of (2). It is a replacement because it denies the three key features of (2) that a Realist believes any alternative to (2) must explain.

For one thing, the Nominalist account denies that (2) is an example of essential predication. this is because color is no longer a constituent in redness. Furthermore, if A is the essence of B, then if B loses A, B ceases to exist.[22] If Campbell loses humanity, then Campbell ceases to exist. Similarly, if redness loses color, redness would cease to exist. But if the Nominalist account of (2d) is correct, (2d) expresses set membership, and the elements of a set do not change when the set changes. If one has a set of the natural numbers from 1 to 10, and then forms a set of natural numbers from 1 to 9, nothing happens to 1. Likewise, if there were no green tropes in the world, then the color set of all sets whose members are color tropes would be different than it is now. But the set of red tropes would not change, nor would each member of that set. Red tropes remain intact regardless of what happens to other tropes.

So the existence of a red trope and its "nature" do not depend on alterations in the set of all sets whose members are color tropes. Thus, the latter is not the essence of the former. Sets are of the essences of their members.

This can be seen in another way. Consider a world where there were no colors other than red. The only color tropes would be red tropes. In this case, redness would be identical to color because two sets are identical just in case their members are identical. "Color" would be just another way of saying "redness." In this world, a red apple would have a red trope but it would not have a color trope as a distinct entity from the red trope. On the other hand, a Realist could argue that when a red apple instances redness in this world, it does instance the universal, color, as a different (though inseparable) entity. If x instances a determinate universal, F, it also instances every determinable universal that is a constituent of F. But these universals are not identical. Sentences (2d) and (2e), especially as Armstrong states them, do not capture this fact. So they are inadequate from a Realist point of view.

Second, the Nominalist account does not ground the membership of the set of red tropes in the set of colors. If it did, then it would make an implicit appeal to a universal that is in each member of that set. And, third, reds are not complex entities but simples in a Nominalist ontology. So, if these three features are part of a correct Realist account of (2), then Nominalist paraphrases are really replacements that fail to do justice to a range of phenomena which motivates Realism in the first place.

It is always open to a Nominalist to argue that this Realist account of (2) is not what is involved in holding that "Red is a color" is true. So, the force of this objection comes down to whether or not the three issues described above are what are really involved in red being a color and whether Nominalist alternatives can really say all that needs to be said about (2).

Even if the Realist uses some model to explain sentences like (2) other than the one I have been employing, I believe he still can argue against a Nominalist paraphrase of (2). Suppose a Realist adopts what some call a determinable/determinate view of (2). According to this view, color is a determinable that covers the nature of and is particularized in the determinate, redness. The first order universal, redness, is color particularized at the level of the first order universal. The generic universal, color, is in some sense identical to the specific universal red, and in some sense different from it. As Brand Blanshard put it: "The universal is thus in it differentiations; it is identical to them; it is distinct from them."[23] The genus is realized in the species and the species is a realization of the genus. The species contains no entity outside the nature of the genus. Rather, the species comes from within it. To be red is to be color, and to be color is to be red or orange or some other determinate shade of color. So according to this view, red is a simple entity and as such, it is a way of being color.

The determinable/determinate view is not my view. But I still believe it or other Realist accounts of higher order predication can raise an objection to a Nominalist account of (2). The Realist could argue that the determinable/determinate relation is one that obtains between universals of different orders but not between a universal and a particular, i.e. a trope. A trope is not a way of being red like redness is a way of being color. Redness is a modification of color, but it is still something general. On the other hand, the subjects of predication in (2d) or (2e), when we say a red is a color, are not a way of being anything, much less color. The connection between a higher and lower order universal is a necessary connection, but the relation between color and a particular red trope is a contingent connection. A red trope is a particular-entity-here, and it is not sufficiently general to be a way of being color.

This point can be seen in another way. A red trope, in some sense, is a-nature-and-a-location. But location is outside the nature of redness. It does not come from within it as a particularization or realization of that nature. There does not seem to be an intimate connection between this location here (where red_1 is located) and the redness of red_1. However, it is easier to see how one could construe redness as somehow "in" color and as the realization of color. Redness seems to be more closely related to color than location is to redness. So it is easier to see how redness is a way of being color than it is to see how this location is a way of being red.

In short, the determinable/determinate structure is neither extendable to the relation between red_1 and its nature nor to the relation between a red and a color. In neither case do we have the subject as a way of being the predicate. So a Nominalist account of predication in general, and (2) in particular, cannot utilize the determinable/determinate relation. But if this relation is required to make sense of (2), then the Nominalist paraphrase in (2d) or (2e) is an inadequate replacement for (2). They do not express the same thing.

The Nominalist could respond by saying that the Realist has begged the question here by already assuming that there is a categorical distinction between universality and particularity in order to make her case that the determinable/determinate relation is a relevant feature of predication between orders of universals. But since there are no universals, this relation is irrelevant and it is a virtue that (2d) or (2e) does not capture it.

3. *An Argument from Husserl*

I wish to raise one final objection against (2d) and (2e). These sentences are general and make claims about a totality of individual reds. On the other hand, the Realist holds that (2) is a singular proposition making a claim about a particular entity, the universal redness. The Realist expresses this conviction in (2a). Thus, the Realist believes that (2d) and (2e) are not adequate paraphrases of (2) and this can be pointed out by focusing on the differences between (2d) and (2e) *vis a vis* (2) and by mentioning some problems with the former.

This line of argument is essentially the one used by Husserl in Sections 1-4 of Chapter 1 of *Logical Investigations II*. Husserl's main point is that we are conscious of universal objects in acts that are different from those in which we are conscious of individual objects, even when those objects are particularized attributes. Thus, the ideal unity of the former cannot be reduced to the dispersed multiplicity of the latter. These acts have different intentional objects. This can be illustrated by three features that highlight the difference.[24]

First, when one intends a group of reds, either at a single glance or in single acts of comparing each red trope to another, there is an implicit recognition of multiplicity and an act of comparison. One notices exact similarities. When one intends the universal redness, no such multiplicity is involved nor is there a need for comparison. We intend the single entity, redness, in its unity. Thus, the two acts are essentially different because their objects are different. It should be noted, as well, that in the second intention the object is redness, a color, and so forth. It is not a set.

Second, I never have before me all the red tropes there are. So if redness is the totality of all reds, then there is more I can learn about redness. If there is a red trope on the surface of Mars, then my knowledge of redness is incomplete until I know about this red trope.[25] On the other hand, when I attend to the universal redness, I know all of it because it is here before me. There is nothing essential to it that is left out for me to know about redness which can be known merely by seeing other reds in the same way I see this one.

Third, when redness is construed as a totality of reds as in (2d) and (2e), the question arises as to what unifies this totality. This question does not even arise, however, when one's intentional object is the universal redness. For it is not a totality.

These three features illustrate that the two acts (or to put the point linguistically, the sentences (2d) and (2e) *vis a vis* (2)) are different because their intentional objects (referents) are different. So, talk about universal objects cannot be accurately reduced to or paraphrased in terms of talk about a dispersed multiplicity of tropes. Husserl provides a fitting summary to this argument: "These distinctions and others like them are quite irremovable. We are not merely dealing with abbreviated expressions: we cannot eliminate such differences through any elaboration or circumscription."[26]

Biola University

IV. Summary

The debate among Extreme Nominalists, Nominalists, and Realists regarding the existence and nature of universals involves the phenomenon of abstract reference and sentences like (1)-(3). Most philosophers agree that these sentences provide serious problems for the Extreme Nominalist but not for the Nominalist. The Nominalist can employ two basic strategies in treating these sentences. He can hold that they contain abstract singular terms referring to sets of tropes. And he can paraphrase these sentences in such a way that they become general claims about a totality of individual tropes. I have argued that both strategies fail. If I am right, then sentences like (1)-(3) provide good reasons for rejecting Nominalism as well as Extreme Nominalism. Opinions to the contrary are simply mistaken.

Received September 5, 1989

NOTES

1. Realists differ over the account they give of higher order predication. Some would hold, for example, that (2) places red in a quality-order. Other that it expresses some sort of determinable/determinate relation. These differences will be discussed later, but for now, they do not matter. All these Realist views agree that (1)-(3) involve reference to universals, and that is the main issue in the debate with Extreme Nominalists and Nominalists.

2. Not all Realists agree with me on this point. In general, Realists often differ over what set of phenomena should provide the major support for Realism, e.g. resemblance for Panayot Butchvarov, predication for Nicholas Wolterstorff, and abstract reference for Michael Loux. I see no reason to emphasize any of these at the expense of the others. For more on this, see J. P. Moreland, *Universals, Qualities, and Quality-Instances* (Lanham: University Press of America, 1985).

3. See D. C. Williams, "On the Elements of Being: I," *The Review of Metaphysics*, vol. 7 (1953), pp. 03-18; "The Elements of Being: II," *The Review of Metaphysics*, vol. 7 (1953), pp. 171-92; G. F. Stout, "The Nature of Universals and Propositions," reprinted in *The Problem of Universals*, ed. by Charles Landesman (New York: Basic Books, 1971), pp. 153-66; Keith Campbell, "Abstract Particulars and the Philosophy of Mind," *Australasian Journal of Philosophy*, vol. 61 (1983), pp. 129-41; "The Metaphysics of Abstract Particulars," in *Midwest Studies in Philosophy Volume VI: The Foundations of Analytic Philosophy*, ed. by Peter A. French, Theodore E. Uehling, and Howard K. Wettstein (Minneapolis: University of Minnesota Press, 1981), pp. 477-88; *Metaphysics: An Introduction* (Encino: Dickenson Publishing Co., 1976). Some classify Husserl as a Nominalist, but I have argued against this elsewhere. See J. P. Moreland, "Was Husserl a Nominalist?" *Philosophy and Phenomenological Research*, vol. 49 (1989), pp. 661-74.

4. For a criticism of Campbell's version of Nominalism, see J. P. Moreland, "Keith Campbell and the Trope View of Predication," *Australasian Journal of Philosophy* (forthcoming).

5. D. M. Armstrong, *Universals and Scientific Realism*, 2 vols. (Cambridge: Cambridge University Press, 1978); Edwin B. Allaire, "Existence, Independence, and Universals," in *Iowa Publications in Philosophy* Vol. I: *Essays in Ontology*, ed. by Edwin B. Allaire (The Hague: Martinus Nijhoff, 1963), pp. 03-13; Gustav Bergmann, *Realism: A Critique of Brentano and Meinong* (Madison: The University of Wisconsin Press, 1967).

6. D. C. Williams, "On the Elements of Being: I," p. 10.

7. Nicholas Wolterstorff, *On Universals* (Chicago: University of Chicago Press, 1970), p. 199.

8. Cf. D. M. Armstrong, *Universals and Scientific Realism*, Vol. I, pp. 58-63; Michael Loux, *Substance and Attribute* (London: D. Reidel, 1978), p. 75. Loux' treatment of sentences like (3) is, in my opinion, inconsistent with Realism. For a defense of this claim, see *Moreland, Universals, Qualities, and Quality-Instances*, pp. 144-58.

9. Cf. Frank Jackson, "Statements About Universals," *Mind*, vol. 86 (1977), pp. 427-29; Armstrong, *Universals and Scientific Realism*, Vol. I, pp. 58-60; Arthur Pap, "Nominalism, Empiricism, and Universals: I," *The Philosophical Quarterly*, vol. 9 (1959), pp. 330-40.

10. I am not assuming that redness is a simple entity for a Realist. For a helpful Realist treatment of the simplicity of colors, see Panayot Butchvarov, *Resemblance and Identity* (Bloomington: Indiana University Press, 1966), pp. 36-47.

11. Armstrong, *Universals and Scientific Realism*, Vol. I, pp. 59-60.

12. *Ibid.*

13. Cf. D. C. Williams, "On the Elements of Being: I," pp. 9-12, for an example of this.

14. Indeed, the very idea of a natural set of resembling tropes is troublesome. For if a trope's nature is identical to its location, then it is difficult to avoid the conclusion that tropes are, after all, bare particulars identical to places. See Moreland, "Keith Campbell and the Trope View of Predication."

15. (1e) also fails because it spells out resemblance in terms of set membership. But the Trope Nominalist wants to explain set membership in natural sets in terms of resemblance. So (1e) would clearly be circular in this case.

16. Armstrong, *Universals and Scientific Realism*, Vol. I, p. 61.

17. *Ibid.*

18. For a very helpful discussion of different Realist views of resemblance between universals and higher order predication as expressed in (2), see Armstrong, *Universals and Scientific Realism*, Vol. II, pp. 101-31.

19. Cf. Michael Loux, "Form Species, and Predication in Metaphysics Z, H, and Θ," *Mind*, vol. 88 (1979), pp. 01-23.

20. For a fuller statement and defense of the view I am presenting here, see Evan Fales, "Generic Universals," *Australasian Journal of Philosophy*, vol. 60 (1982), pp. 29-39.

21. J. R. Jones, "What Do We Mean by An 'Instance'?," *Analysis*, vol. 11 (1950), pp. 11-18.

22. By viewing the relation between higher and lower orders of universals as a relation of essential predication, we have an explanation for why higher order universals transcend lower ones. Color can exist without redness but not vice versa because color is an essential constituent in redness, but redness is not an essential constituent in color.

23. Brand Blanshard, *The Nature of Thought*, 2 vols. (London: George Allen and Unwin Ltd., 1939), Vol. I, p. 611.

24. Husserl's argument focuses on mental acts, their character, and their intentional objects. He does not focus on words or sentences. But this does not affect the application of his points to (2), (2d), and (2e), since it is primarily the use and not the mention of these sentences that is important here.

25. This point was suggested to me by Dallas Willard.

26. Edmund Husserl, *Logical Investigations*, 2 vols., tr. by J. N. Findlay (London: Routledge & Kegan Paul, 1970), Vol. I, p. 341. It could be argued that Husserl's points about having an object like redness directly before me and attending to it immediately is hopelessly old-fashioned, given that Wittgenstein, Kripke, Goodman and others have shown that thought is rule governed activity. Thus, we have no idea of what having an object "here directly before me" would mean. The issues involved in this objection are too complex to be dealt with here (Is there a private language? Do we need to think by means of language at all? Is perceptual realism true? and so on). If one sides with the early Husserl on these questions, then this objection loses its force. Further, Husserl's arguments could be recast in terms of linguistic rules and sentences. Thus, the rules involved in understanding and applying sentences like (2) utilize subject terms which prima facie behave like abstract singular terms. On the other hand, sentences like (2d) and (2e) involve rules which make reference to the notions of multiplicity and totalities (e.g., classes and sub-classes), diversity/similarity in respects of resemblance between or among objects, and the dispersal of the objects referred to by the terms used in those sentences (e.g., "particular colorednesses"). Thus, Husserl's point can be made in terms of linguistic entities and rules.

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 27, Number 4, October 1990

WHAT IS VIRTUE ETHICS ALL ABOUT?

Gregory Trianosky

THE past fifteen years have witnessed a dramatic resurgence of philosophical interest in the virtues. The charge that modern philosophical thought neglects the virtues (Becker 1975, Von Wright 1963, Taylor in French, Wettstein, and Uehling 1988), once apposite, is by now outmoded; and the calls for a renewed investigation of virtue and virtue ethics are being answered from many quarters. What has been missing to date is any systematic guide to the plethora of issues, charges, claims, and counter-claims raised in recent work on the virtues. This survey takes the first steps toward charting this vast and vastly exciting terrain.[1]

Interestingly, not all those who are engaged in the new investigations of virtue agree in endorsing an *ethics of virtue*, or for that matter any single substantive position. Instead concentration on the virtues has served as a rallying point for many writers opposed in different ways to the main tendencies of post-eighteenth-century thinking in ethics. In particular, what unifies recent work on the virtues is its opposition to various central elements of a view which I will call *neo-Kantianism* (Cf. Blum 1980, pp. 1-3). This is not necessarily Kant's own view, as a number of able commentators have pointed out recently, although its elements are in their ancestry recognizably Kantian. Nor is it necessarily a view held in *toto* by any one contemporary moral philosopher. (See Donagan 1977; Darwall 1983; Gewirth 1978 for defenses of important elements of neo-Kantianism.) Nor, finally, is it always a view whose component claims are either uniformly understood or carefully distinguished by its adversaries. Its tenets follow.

(1) The most important question in morality is, "what is it right or obligatory to do?"

(2) Basic moral judgments are judgments about the rightness of actions.

(3) Basic moral judgments take the form of general rules or principles of right action. Particular judgments of the right are always instances of these.

(4) Basic moral judgments are universal in form. They contain no essential reference to particular persons or particular relationships in which the agent may stand.

(5) Basic moral judgments *are not* grounded on some account of the human good which is itself entirely independent of morality.

(6) Basic moral judgments are categorical imperatives. They have a certain "automatic reason-giving [justificatory] force" (Foot 1978, p. 161) independently of their relation to the desires and/or interests of the agent.

(7) It is possible for considerations about what is required by basic moral judgments to play some role in the actual motivation *of any agent*, independently of the operation of desire and emotion in him/her.

(8) It is necessary that considerations about what is required by basic moral judgments play some role in the actual motivation of *the truly virtuous agent*, independently of the operation of desire and emotion in him/her.

(9) The virtuousness of a trait is always derivative from some relationship it displays to what is antecedently specified as right action.

Nearly all contemporary writers on the virtues do agree in rejecting (1). Their work always shows and typically says that the emphasis in moral philosophy should shift from investigations of the right to investigations of virtue, the virtues, and the virtuous life (Taylor in French, Wettstein, and Uehling 1988; Becker 1975; Anscombe 1958). Further, nearly every contemporary writer on the virtues rejects at least one more of these nine claims. Our study of recent work on the virtues will be the study of how the assaults on each of these nine claims are interrelated, both logically and dialectically.

I

We may begin by considering one central contrast, that between the ethics of duty and the ethics of virtue. To speak roughly, in its pure form an ethics of duty holds that only judgments about right action are basic in morality, and that the virtuousness of traits is always derivative in some way from the prior rightness of actions (see e.g. Gewirth 1985, p. 751). Conversely, an ethics of virtue in its pure form holds that only judgments about virtue are basic in morality, and that the rightness of actions is always somehow derivative from the virtuousness of traits. The conjunction of neo-Kantian claims (2) and (9) constitutes an endorsement of one form of the pure ethics of duty and a rejection of its contrary, the pure ethics of virtue. The perceptual intuitionism endorsed by Prichard and Ross' Aristotle constitutes the endorsement of another form of the pure ethics of duty.

Formulated more precisely, a pure ethics of virtue makes two claims. First it claims that at least some judgments about virtue can be validated independently of any appeal to judgments about the rightness of actions. In Plato's *Republic* for example it appears to be simply the harmonious order of the just person's psyche which makes it good, and not, say, its aptness to produce right action.

Second, according to a pure ethics of virtue it is this antecedent goodness of traits which ultimately makes any right act right. For instance, Plato says that just actions are those which produce and maintain that harmonious condition of the psyche (*Republic* 443e); and Aristotle might be read as saying that what one ought to do is what the virtuous person, or the person of practical wisdom, would do. In both these cases the rightness of action supervenes on some appropriate relation to what is antecedently established as virtue.

Obviously views about exactly what is to count as "an appropriate relation" to virtue will vary. (Our discussion below in section IV of the varieties of the ethics of duty should suggest by analogy some of the forms such views may take.) But in any case for the pure ethic of virtue the moral goodness of traits is always both independent of the rightness of actions and in some way originative of it as well. The same points may be made, *mutatis mutandis*, about the pure ethics of duty.

II

The debate which inaugurated much of the renewed interest in the virtues began in Anscombe's well-known article, "Modern Moral Philosophy" (1958). Anscombe challenges claim (6). She finds the notion of a universal moral law which is not the command of any deity, a "special moral 'ought,'" unintelligible (cf. Taylor 1985). It might remain to "look for norms," she says, which are grounded in the facts about what we need in order to "flourish." And perhaps what it is for us to flourish, she suggests, is as Aristotle held to live a life informed by virtue.

The rejection of (6) is often equated with the embracing of this Aristotelian virtue ethic. In point of fact, linking them requires several controversial steps. Anscombe's own argument is a case in point. First she claims that there is no secularized moral "ought" which has an intelligible application to all rational beings, or even to all human beings, independently of their interests and desires. Second, she claims that there is an "ordinary" ought or norm, which applies in some version or other to every living creature. This is the "ought" which instructs us about what is good for us. Third, there is the claim that the notion of our good is to be parsed in terms of what we "need" or require in order to flourish. Finally, there is the characteristically Aristotelian claim that "the flourishing of a man *qua* man consists in his being good...a man needs, or ought to perform, only virtuous actions." Plainly one might reject (6), and endorse the first two of these claims, and yet reject the third and/or fourth. The movement to the Aristotelian ethic of virtue is a natural dialectical progression from the rejection of (6). But it is not a logical consequence of it, taken alone.

Foot's well-known "Morality as a System of Hypothetical Imperatives" (1978) has also been influential in the rejection of (1) and (6), if not in the development of an ethics of virtue proper. She claims that although moral requirements may have a legitimate *application* to an agent independently of whether his desires and interests are served by conformity, these requirements do not give *reasons to act* which are independent of whether his desires and interests are thus served. Indeed, there are no such reasons (1978, pp. 179, 148-56). Thus if a thoroughly wicked person has no desire or interest which is promoted by his or her being moral, he or she has

no reason to be moral (1978, p. 161). Foot concludes that the reason-giving force of moral requirements is always conditional. They constitute only hypothetical imperatives and not categorical ones.

Here again, no position on the issue of whether basic moral judgments are in Foot's sense categorical or hypothetical entails the rejection of claims (2) and (9), and the acceptance of any version of the ethics of virtue as we have defined it; nor conversely. Yet Foot's discussion reveals that larger questions about moral motivation are connected in a variety of ways to questions about the nature of the reasons there are to be moral.

For one thing, the dispute over (6) is linked with arguments over what sorts of moral motivation are possible or desirable by certain kinds of *internalist* suppositions. That is to say, writers on the virtues often hold (for differing reasons) that philosophical claims about the nature of moral requirement must somehow be modeled in the motivational structure of moral agents. Many neo-Kantians seem to agree. For example, they suppose that since moral requirements are categorical, (7) must be true: it must be at least possible for any moral agent to be motivated by a "sense of duty," conceived as wholly distinct from desire (Darwall 1983). Moreover, a commitment to (7) is characteristically paired with a commitment to (4), the claim that basic moral principles are impartial in their content. (4) and (7) together imply that it is possible for moral agents to be motivated impartially, simply by considerations about what basic moral requirements dictate. Rejection of either the possibility or the worth (moral or otherwise) of such impartial motivation is a recurring theme in the work of contemporary writers on the role of character in ethics (Williams 1981; Stocker in Kruschwitz and Roberts 1987; Blum 1980, pp. 142f; Wolf in Kruschwitz and Roberts 1987). It should be noted, however, that in many cases the arguments of these anti-Kantian writers go through only given much stronger formulations of internalist doctrines like (7) or (8) than the ones presented here. The work of Herman (1981) and Baron (1983, 1984) forcefully suggests that no such stronger formulations are entailed by (6), or by any other central Kantian thesis.

Disputes about the moral worth of various patterns of motivation enter the arena in yet another way in the discussion of (8). Suppose for a moment that we follow Foot and many others in identifying a core group of traits as *substantive virtues*. These are traits which involve a powerful and enduring concern for some morally valuable end like the well-being of others, the telling of the truth, or the keeping of agreements (1978, pp. 165-66, 154-55). Then we might adopt a distinction of Prichard's which has had some influence in the contemporary literature on virtue (Frankena 1970). According to Prichard *moral goodness* involves the disposition to be motivated by a sense of duty, conceived as independent of desire. On the other hand a (substantive) *virtue* is simply a standing, intrinsic desire for some morally significant end, *where the end is described wholly in non-moral terms*. One can then think of the "ethics of virtue" in the way that some claim Foot does (Baron 1983), as holding that the truly moral agent need only display the substantive "virtues," and need not be "morally good." The "ethics of virtue" will then hold up the ideal of a near-preternatural innocence as true virtue (Trianosky 1990). The "ethics of duty," at least in its most well-known form, will be some version of the claim expressed in (8).

I think this use of virtue and duty terminology is infelicitous, but no real harm is done so long as we understand that the ethics of virtue as we initially defined it does not commit one to the "ethics of virtue" as it is understood here. Ethics of virtue advocates could agree that some "practical" element of responsiveness to basic moral principles must enter into the makeup of any truly virtuous agent. Of course they will tend to conceive this element as involving a responsiveness to considerations about (say) what a person of practical wisdom would do, rather than strictly to considerations about duty. This is not really a disagreement over (8), however, but instead over (2), and perhaps its frequent consort (3).

III

Since its introduction into the contemporary discussion by Anscombe, the ethics of virtue has come to be seen as a "third option," competing with both deontological and utilitarian views. On this way of looking at the matter, deontological theories take judgments about the right as basic; utilitarianism takes judgments about "the desirability of certain states of affairs that [actions] produce" as basic; and the ethics of virtue "derives the desirability of the act from the desirability of...motives or traits of character..." (Dent 1984, pp. 32-34; Louden 1984).

This classificatory approach encourages one to

overlook the radical alternatives which the utilitarian tradition can provide to the ethics of duty. To be sure, the most familiar utilitarian views characterize virtues either as dispositions (however complex) to do what would be in fact right by act-utilitarian standards (Hare 1981); or as traits which maximize the probability that we will do what would be right by such standards (cf. Railton 1984, pp. 152-56). But recent work by Brandt (1981) and Adams (1976) describes forms of trait- and motive-utilitarianism in which the relation between moral character and utility is not thus mediated by an account of the right. On one such view, for instance, what makes a trait a virtue is simply that its general possession would maximize utility. No reference to any theory of the right, utilitarian or otherwise, is presupposed by this view of virtue.

I suggest that we may more perspicuously divide ethical theories along two distinct and orthogonal lines. First, they may be divided into the ethics of virtue and the ethics of duty in the manner indicated at the outset. This distinction divides ethical theories on the question of what sort of *moral* judgments they take to be basic. Next, each of these ethics may take either a teleological or a non-teleological, "deontological" form. This second distinction divides ethical theories on the question of whether they take the basic moral judgments themselves to be autonomous or derivative. Pure *teleological* theories hold that basic moral judgments are to be grounded on some account of the good, where the good is conceived as describable *independently of any reference either to moral rightness or to virtue:* pleasure, the satisfaction of desire, or the obtaining of various intrinsically desirable states of affairs, for example. Pure teleological theories deny that ethics is autonomous because they hold that all judgments about moral value, if I may use this as a generic term for rightness and moral virtue, must ultimately be grounded in this way on judgments about *non-moral* value. Non-teleological or "deontological" theories disagree, holding instead that basic moral judgments, whether about virtue or about duty, are not grounded on considerations about the (non-moral) good. In this way pure non-teleological theories hold that moral value is autonomous, or not dependent for its philosophical justification on any claims about non-moral value (Scanlon 1982). (5), one of the central propositions of neo-Kantian ethics, embodies the non-teleological claim, albeit in a comparatively weak form. So, as I will indicate below, do the writings of certain virtue theorists.

(I have used the label "teleological" to advertise that the range of views here is not limited to strictly utilitarian theories. What is characteristic of utilitarianism is its insistence that the rightness of actions or the virtuousness of traits depends on their *causal* relation to the good. But the relation between an action or a trait and the good may be intentional rather than causal. The act or trait may *aim* at human well-being, for example, rather than simply helping to promote it. Neither of these necessarily implies the other. Non-utilitarian teleological theories might measure the rightness of acts or the virtuousness of traits not by their effects but simply by the extent to which they take the good as their intentional object. Thus Frances Hutcheson, for example, famously says that "benevolence is the whole of virtue," because he takes a virtuous person to be one who aims at the good.)

Now it should be clear that the sorts of issues which lie between the ethics of duty and the ethics of virtue as we have characterized them will be quite different from those raised by the "deontology-teleology" debate. Moreover, it follows from this way of understanding things that just as there are both teleological and deontological forms of the ethics of duty, familiar from long-standing debates in this and the previous two centuries, so also there are both teleological and non-teleological or (as it were) "deontological" forms of the ethics of virtue. Many of these last are only now beginning to be explored again.

Non-teleological ethics of virtue are theories which maintain against all forms of utilitarianism that the basic judgments about virtuous traits *can* be grounded without appeal to any independently-formulated account of the good. Like the familiar deontological views in the ethics of duty, these theories hold that basic moral judgments, however conceived, may be justified *autonomously*. The most familiar non-teleological theories are perfectionist ones like Aristotle's, on which virtue is a constitutive element (if not the central element) of the human good rather than merely a means, however indispensable, to its attainment (Larmore 1987, pp. 30-36).

This way of drawing distinctions suggests that the relationship between utilitarianism and contemporary writers on the virtues is likely to be an ambigu-

ous one, and so it is. On the one hand, utilitarians as I have described them plainly reject (5). But given the distinctions I have drawn, no position on (5) will follow from a virtue theorist's rejection of (2) and (9) and the acceptance of a pure ethics of virtue. Indeed, perfectionist versions of the non-teleological ethic of virtue will endorse (5) as it is written.

Nonetheless, one of the greatest and most widespread sources of dissatisfaction among contemporary writers on virtue is the way in which the Kantian tradition does take morality to be autonomous in the extreme, cut off entirely from the human good at its base. Neo-Kantians typically will endorse not only (5) but the stronger thesis that basic moral principles are not to be grounded on any account of the human good at all, whether independent of morality or not (Donagan 1977; Darwall 1983). Advocates of a perfectionist ethic of virtue will join utilitarians in rejecting this stronger thesis; for they all agree against such neo-Kantians that some conception of the human good must figure centrally in our account of moral value. The renewed interest in Aristotle so familiar in the writings of virtue theorists is but one sign of this widely-held conviction.

Plainly, the crucial issue for writers on the virtues who wish to affirm a close relation between virtue and the good is thus whether such a conception of the good can be described independently of reference to moral virtue. Teleological virtue theorists like the motive- and trait-utilitarians say "yes." Non-teleological, perfectionist virtue theorists say "no." What is surprising is how many of the most influential writers on the virtues today in fact defend some sort of teleological if not specifically utilitarian answer. Von Wright's (1963, p. 140) remark is representative: "Virtues...are needed in the service of the good of man. This usefulness of theirs is their meaning and [natural] purpose" (cf. Wallace 1978, Warnock 1971, Geach 1977, MacIntyre 1981).[2]

To be sure, what distinguishes many of these writers most sharply from more brazenly utilitarian writers like Hare or Smart is that their conception of the good is far richer. They are not hedonists, nor satisfaction-of-desire theorists. They speak instead of human flourishing or well-being, or of practices, traditions, and narratives. This enrichment of our philosophical repertoire of conceptions of the good is without question a very important contribution to modern moral philosophy. But utilitarianism is one kind of teleological doctrine about the foundations of ethics. It holds that the right and the virtuous must somehow promote the (non-moral) good. It is not necessarily committed to any particular substantive theory of that good. Hence writers on the virtues cannot avoid a commitment to utilitarianism by offering a more sophisticated theory of the good.

The possibility of a non-teleological ethics of virtue which seek to retain a close relation between virtue and the human good is precisely what Elizabeth Anscombe suggested (1958; cf Geach 1977, pp. 9-12). This suggestion has set the agenda for the future for many writers on the virtues. The difficulty for some of these has been how to adopt some more or less Aristotelian notion of human flourishing without Aristotelian metaphysical commitments, but at the same time without abandoning the search for an alternative to utilitarianism (cf. Wallace 1978, p. 34; MacIntyre 1981, pp. 187-89). Others (Taylor 1985; Nussbaum in French, Uehling, and Wettstein 1988) seem willing simply to endorse Aristotle's claim that our function is a life of rational activity despite whatever metaphysical difficulties may attend it.

Despite the great deal of work that remains to be done on them, non-teleological ethics of virtue offer important advantages over any other view. They do justice to two guiding intuitions which seem at first to be irreconcilably at odds. The first is the minimal Kantian idea, expressed in (5), that morality is autonomous. The second is the idea that, as utilitarians have always insisted, morality is essentially connected with the human good. Defenders of the non-teleological ethics of virtue can accept this latter utilitarian idea, for they can maintain that virtue is a constitutive element of the human good. By the same token, although they reject any extreme neo-Kantian version of (5), they can accept (5) as I have formulated it.

We may close this section by noting that here we have discovered one final connection between a non-teleological ethic of virtue and the debate over reasons for action which was the focus of (6). One may combine a perfectionist view of the human good with the claim made by both Anscombe and Foot that one always has reason to do what is in one's interest. Taken together these entail that moral virtue is a form of *human excellence*, or a state of character which there is reason to pursue for its own sake as a constituent of one's good. I believe this is the position that Anscombe was suggesting and that

Foot is working toward. It is certainly a position which any admirer of Aristotle's ethics should take seriously.

IV

I remarked at the outset that not all writers on the virtues defend the ethics of virtue as I have defined it. Indeed, when one comes to examine prevalent views about what the virtues are and what makes them virtues, it becomes clear that many of these writers still operate within the framework of an ethic of duty. Seeing how this is so will demonstrate some of both the merits and the limitations of recent work on the virtues.

Earlier we made a distinction between two kinds of teleological theories: causal or utilitarian theories and intentional theories. They differed in their account of what the proper relation was between morality and the (non-moral) good. This distinction cut across the ethics of duty/ethics of virtue distinction with which we began. Here we may make yet another similar distinction within the ethics of duty, having to do with the view of the virtues taken by such an ethic. Despite the similarities of structure, this new distinction will be orthogonal to both the previous distinctions.

Recall that the ethics of duty took judgments about the right to be fundamental in morality. Moreover, it held that the virtuousness of traits was always wholly derivative, in one way or another, from the prior rightness of actions.

Whether its criterion of rightness happens to be deontological or teleological, therefore, we might say that an ethic of duty always affirms that virtue consists in *a proper orientation toward the right.* But this notion of proper orientation toward the right typically has been interpreted in one of two distinct ways. On the one hand, it has often been understood as involving a disposition to *choose for the sake of* what is antecedently established as right. (Gert (1981) and Brandt (1981) offer quite different perspectives on what it is to have such a disposition.) The radical Kantian view suggested by certain familiar passages in the *Groundwork*, that only the concern to do what is right *as such* constitutes true virtue, is a limiting case of the general idea that all the virtues are substantive concerns of this sort. A surprising number of contemporary authors seem to agree with this general idea (Rawls 1970; Brandt in French, Wettstein, and Uehling 1988; Warnock 1971; Gert 1981), even if they do not agree with Kant's specific contention that all the virtues are varieties of the sense of duty. (Cf. Wallace 1978 on the virtues of conscientiousness *versus* the virtues of benevolence.) These writers thus all offer an account of the virtues which is entirely compatible with the ethics of duty.

On the other hand, the notion of being properly oriented toward the right has been interpreted in causal terms. The virtues have then been understood as whatever traits generally serve to enable human beings *better to pursue their commitment to* what is antecedently identified as right. So understood the virtues may be quite a heterogeneous lot, or they may all be traits of the same kind (Von Wright 1963, Roberts in Kruschwitz and Roberts 1987). But to conceive the virtues as enablers of this sort is again to deny no tenet of the ethics of duty. Moreover, as I have suggested, within the ethics of duty both deontologists and teleologists may endorse either of these conceptions of virtue.

Certain current writers on the virtues hold the moderately pluralistic view that some virtues are substantive dispositions to choose what is right, while the rest are traits which enable right action (Foot 1978, pp. 1-18; Hare 1981). Here again, of course, this moderate pluralism remains entirely consistent with the central ethics of duty claim, that the moral worth of virtuous traits is wholly derivative from the rightness of the actions to which they are somehow related.

One of our earlier distinctions can now be shown to be extensionally equivalent to a version of this same substantive virtue/enabling virtue dichotomy, at least within the comparatively restricted domain of the *teleological ethics of virtue.* Now a teleological ethic of virtue conceives of the virtues as consisting in proper orientation toward *the* (non-moral) *good*, rather than toward the right. But recall the distinction in section III between two broad teleological conceptions of the relation between morality and that good: On the one hand morality may be understood as involving some suitable *intentional* orientation toward the good; or on the other hand it may be understood as involving a suitable *causal* orientation toward the good. If one defends a teleological ethics of virtue, then in the former case the virtues may again be interpreted as involving substantive dispositions to choose for the sake of what

is good. In the latter case, correspondingly, they may be interpreted as traits which generally enable the pursuit of the good. Indeed, teleological virtue theorists have generally conceived of the virtues on one of these two models; although naturally teleological duty theorists have not (Adams 1976). Our earlier distinction between two types of teleological theories is thus extensionally equivalent to a version of the distinction between the conception of virtue as substantive and the conception of virtue as enabling, but only within the realm of teleological virtue ethics. Within the realm of teleological duty ethics the two distinctions remain orthogonal.

A less modest and even more reasonable pluralism about the virtues would amalgamate all these various conceptions of virtue, and hold the mixed ethics of duty/ethics of virtue view that virtues must all be species of proper orientation toward *either* the right *or* some independently conceived non-moral good; and that such a proper orientation could in each case be interpreted as involving either a substantive disposition to choose or some enabling trait (W. D. Ross 1930, p. 134).

However, even this more reasonable pluralism commits one to rejecting a non-teleological ethic of virtue, in which autonomous judgments of virtue are fundamental. Hence it makes it difficult, though not impossible, to accept the Aristotelian position I mentioned at the end of section III, on which virtues are conceived as human excellences (Frankena 1980).

It is also too modest in a more intramural respect. For one thing, there may be a variety of significant causal relationships between traits and the good or the right besides the enabling-relation between trait and action. (One's wisdom may win the respect of others and so make them more responsive to one's advice, for example; and one's willingness to abide by one's agreements may allow others in the community better to pursue their ends.) Analogously, there are many intentional states which are not fundamentally dispositions to behave, to choose, or to deliberate. Yet surely one's attitudes and emotions about the right or the good, for example, may be either appropriate or inappropriate. One may venerate the moral law, or one may obey its dictates resentfully. One may love the good wholeheartedly, or one may take a secret pleasure in what is evil. A full appreciation of the richness of the intuitive idea that virtue consists in proper orientation toward the right and/or the good thus awaits further exploration of the great variety of intentional states which may take the right or the good—or for that matter the bad or the wrong—as their objects (Dent 1985, p. 29; Trianosky 1988); as well as the variety of causal relations which may obtain between traits and moral value.

V

Perhaps the most frequently-heard objection to virtue ethics is that "[it] is structurally unable to say much of anything" about what people ought to do. "What can a virtues and vices approach say about specific moral dilemmas," critics ask (Louden in Kruschwitz and Roberts 1987). The real force of this objection is properly directed against a pure ethics of virtue, or perhaps against a position which proposes to substitute a theory of virtue for a theory of right action. Much of what is being written about the virtues involves an endorsement neither of such a pure position nor of such a substitution; and there is certainly no denying that writers on the virtues have begun to take great interest in concrete moral issues (e.g. Foot 1978; Hursthouse 1987). Moreover, here as in section II some form of internalist supposition is operating. Here it seems to be the assumption that any theory of the virtue-making characteristics of traits and actions must at the same time constitute an account of valid moral decision procedures. In general, however, this sort of internalist supposition about theories of moral value seems questionable (Smith 1988, Railton 1984); and typically no argument is given for why virtue theorists in particular must be committed to it.

Perhaps not surprisingly, Pincoffs (1986) and many other virtue theorists seem implicitly to accept the relevant internalist supposition. They handle the objection by instead rejecting the supposition that the central business of moral theories is to help us resolve "moral quandaries." But even if resolving moral quandaries is not the premier task of an ethical theory, there still remains the question of what the ethics of virtue can say about that aspect of the moral life which does involve choosing courses of action.

One answer to this question seeks to reject claim (3), replacing rules and principles with a different sort of standard for right action. The suggestion is that some personal moral ideal is to guide one's decisionmaking (Cua 1978; Pincoffs 1986, p. 24,

Moravscik 1980). Personal moral ideals are understood here as articulating some conception of virtuous character. Interestingly, however, the advocates of this suggestion usually conceive ideals as operating subject to constraints imposed by "the big rules," as Moravscik calls them (1980, pp. 219-20), or "the minimal requirements" of universalizability, as Pincoffs calls them (1986, p. 35). Even in MacIntyre's argument for a complete reorientation of our understanding of ourselves and our lives toward the virtues, there is a "morality of law" which is not just coordinate with the "morality of virtue," but which seems to impose basic constraints on how decisions are to be made (1981, pp. 141-43, 187). Hence this answer seems entirely compatible with a commitment to the pure ethics of duty.

A more radical *tu quoque* response against the ethics of duty is that no theory of the right can constitute a complete guide to action without being supplemented by a theory of virtue. This response has been defended on several levels. First, it has sometimes been pointed out that rules or principles of right action must be applied, and conflicts between them adjudicated. But the rules themselves do not tell us how to apply them in specific situations, let alone how to apply them well, or indeed when to excuse people for failing to comply with them. For these tasks, it is claimed, an account of the virtues is required (Cua 1978; Becker 1975; Moravscik 1980).

Next, it has been argued that much of right conduct cannot be codified in rules or principles (McDowell 1979). Moral situations are too complex; moral rules too general and simplistic. Instead, various substantive virtues like benevolence, honesty, justice, and generosity may each be seen as embodying a form of responsiveness to some range of moral considerations about helping others, truth-telling, and so on. Enabling virtues like sensitivity and empathy operate to ensure that one does not overlook relevant considerations in the particular case. On this model there will be no fixed rules or decision procedures which tell one how to weigh these competing considerations once they are identified (Pincoffs 1986). Moreover, except when one can look to some morally exemplary or paradigmatic individual (Cua 1978), the extent to which one decides well will depend largely on the extent to which one has already developed a virtuous character.

The University of Michigan-Flint

What makes an act right in these instances will instead be simply that it is the particular choice endorsed by thoughtful judgment or practical wisdom, informed by virtuous concern. In these instances at least judgments of virtue will be primary and judgments of rightness derivative. (Whether it follows that the pure ethics of duty defended in claim (9) is mistaken depends on whether these judgments of virtue in turn are basic, or whether they are themselves derivative from even more fundamental principles of the right. For example, it is consistent with the conclusion of the *tu quoque* argument as I have formulated it that a complete account of the nature of practical wisdom in turn requires essential reference to some fundamental principles of the right. Hence it is consistent with that conclusion that both the ethics of duty and the ethics of virtue, taken on their own, are incomplete as action guides.)

In any case it certainly follows that insofar as an ethics of duty is committed to claim (3), it must be incomplete as an action guide. A significant portion of the account of right action will be formulable only in the language of virtue, and not in the language of rules or principles of duty.

It is perhaps out of a desire to show that the judgments of virtue involved here are indeed basic and not derivative from other principles of the right that writers on the virtues characteristically resist the sometimes-suggested translation of statements about virtue into the idiom of rules of duty (Warnock 1971). They maintain that typically there is no codifiable rule or principle of the right which covers just that set of actions characteristic of a given virtue. That set will be describable solely by reference to the virtue itself. (Blum 1980, p. 142; Von Wright 1963; Burnyeat 1971; Dent 1985, pp. 29-30; Pincoffs 1986, pp. 77-78. But see Trianosky 1987.)

If the *tu quoque* argument succeeds then there is at least one powerful reason to study the virtues. But whether it succeeds or not, perhaps the most persuasive argument in favor of studying the virtues is simply that they are the stuff of which much of the moralities of everyday life are made (Sabini and Silver 1982; Thomas 1989). If we are to give moral experience precedence over moral theorizing, we must study the rich and subtle phenomena of moral character.

Received June 26, 1989

NOTES

1. (Pence 1984) is the only other survey of recent work on the virtues of which I am aware. It is less systematic and more oriented toward thumbnail sketches of the canonical texts. Because it is readily available, I have given less space to summary and more space to systematizing the issues addressed by writers on the virtues. Kruschwitz and Roberts (1987) contains a fairly comprehensive bibliography of recent work on the virtues, including work on particular virtues like generosity, courage, humility, and mercy. Card and Hunt (1991) includes a large number of excellent articles on particular virtues.

2. Perfectionist virtue theorists will reject even MacIntyre's sophisticated utilitarian conception of virtues like truthfulness, justice, and courage: "[The virtues are] those dispositions which...sustain us in the relevant quest" for the good, by enabling us to overcome the harms, dangers, temptations and distractions which we encounter. "...The good life for man is the life spent in seeking for the good life for man, and the virtues necessary for the seeking are those which will enable us to understand what more and what else the good life for man is" (1981, p. 204).

3. An earlier and much longer version of this paper was presented to MAPPS, the newly-formed Orange County Moral and Political Philosophy Society, in December 1988. I am very grateful for the comments of Society members, especially David Estlund, Craig Ihara, Alan Nelson, and Gary Watson. I am also very grateful for the comments of Rachel Cohon, Amy Gutman, Brad Hooker, Joel Kupperman, Michael Slote, and a number of others.

REFERENCES

Adams, Robert M. (1976) "Motive-Utilitarianism," *Journal of Philosophy*, vol. 73, pp. 467-81.
Anscombe, G. E. M. (1958) "Modern Moral Philosophy," *Philosophy*, vol. 33, pp. 1-19.
Baron, Marcia (1984) "The Alleged Moral Repugnance of Acting from Duty," *Journal of Philosophy*, vol. 81, pp. 197-220.
Baron, Marcia (1983) "On De-Kantianizing the Perfectly Moral Person," *Journal of Value Inquiry*, vol. 17, pp. 281-93.
Becker, Lawrence (1975) "The Neglect of Virtue," *Ethics*, vol. 85, pp. 110-22.
Blum, Lawrence (1980) *Friendship, Altruism, and Morality* (London: Routledge and Kegan Paul).
Brandt, Richard B. (1981) "W. K. Frankena and Ethics of Virtue," *The Monist*, vol. 64, pp. 271-92.
Burnyeat, M. F. (1971) "Virtues in Action" in *The Philosophy of Socrates*, Gregory Vlastos (ed.) (New York: Doubleday Anchor Books).
Card, Claudia and Hunt, Lester (eds.) (1991) *Character: Essays in Moral Psychology* (Ithaca: Cornell University Press).
Cua, Antonio (1978) *Dimensions of Moral Creativity* (University Park: Pennsylvania State University Press).
Darwall, Stephen L. (1986) "Agent-Centered Restrictions From the Inside Out," *Philosophical Studies*, vol. 50, pp. 291-319.
Darwall, Stephen L. (1983) *Impartial Reason* (Ithaca: Cornell University Press).
Dent, N. J. H. (1984) *The Moral Psychology of the Virtues* (Cambridge: Cambridge University Press).
Donagan, Alan (1977) *The Theory of Morality* (Chicago: University of Chicago Press).
Flanagan, Owen and Rorty, Amelie (eds.) (1990) *Identity, Character, and Morality* (Cambridge: MIT Press).
Foot, Philippa (1978) *Virtues and Vices* (Berkeley: University of California Press).
Frankena, William K. (1980) "The Carus Lectures of William Frankena: Three Questions About Morality," *Monist*, vol. 63, pp. 3-47.
Frankena, William K. (1970) "Prichard and the Ethics of Virtue," *Monist*, vol. 54, pp. 01-17.
French, Peter A.; Uehling, Theodore E.; and Wettstein, Howard K. (1988) *Midwest Studies in Philosophy*, Vol. XIII, *Character and Virtue* (Notre Dame: University of Notre Dame Press).
Geach, Peter (1977) *The Virtues* (Cambridge: Cambridge University Press).
Gert, Bernard (1981) *The Moral Rules* (New York: Harper Torchbook).
Gewirth, Alan (1985) "Rights and Virtues," *Review of Metaphysics*, vol. 38, pp. 739-62.
Gewirth, Alan (1978) *Reason and Morality* (Chicago: University of Chicago Press).
Hare, Richard M. (1981) *Moral Thinking* (Oxford: Oxford University Press).
Herman, Barbara (1981) "On the Value of Acting from the Motive of Duty," *Philosophical Review*, vol. 90, pp. 359-82.
Hudson, Stephen D. (1985) *Human Character and Morality* (London: Routledge & Kegan Paul).
Hursthouse, Rosalind (1987) *Beginning Lives* (Oxford: Basil Blackwell Ltd.).
Kruschwitz, Robert B. and Roberts, Robert C. (eds.) (1987) *The Virtues: Contemporary Essays in Moral Character* (Belmont, California: Wadsworth Publishing).

Larmore, Charles E. (1987) *Patterns of Moral Complexity* (Cambridge: Cambridge University Press).
Louden, Robert B. (1986) "Kant's Virtue Ethics," *Philosophy*, vol. 61, pp. 473-89.
Louden, Robert B. (1984) "On Some Vices of Virtue Ethics," *American Philosophical Quarterly*, vol. 21, pp. 227-36.
MacIntyre, Alasdair (1981) *After Virtue* (Notre Dame: University of Notre Dame Press).
McDowell, John (1979) "Virtue and Reason," *Monist*, vol. 62, pp. 331-50.
Moravscik, Julius (1980) "On What We Aim At and How We Live," in David J. Depew (ed.), *The Greeks and The Good Life* (Fullerton: Department of Philosophy, California State University).
O'Neill, Onora (1983) "Kant After Virtue," *Inquiry*, vol. 26, pp. 387-405.
Pence, Gregory E. (1984) "Recent Work on Virtues," *American Philosophical Quarterly*, vol. 21, pp. 281-97.
Pincoffs, Edmund (1986) *Quandaries and Virtues* (Lawrence: University Press of Kansas).
Railton, Peter (1984) "Alienation, Consequentialism, and the Demands of Morality," *Philosophy and Public Affairs*, vol. 13, pp. 134-71.
Rawls, John (1970) *A Theory of Justice* (Cambridge: Harvard University Press).
Ross, W. D. (1930) *The Right and The Good* (Oxford: Oxford University Press).
Sabini, John and Maury Silver (1982) *Moralities of Everyday Life* (Oxford: Oxford University Press).
Scanlon, Thomas M. (1982) "Contractarianism and Utilitarianism," in Amartya Sen and Bernard Williams (eds.), *Utilitarianism and Beyond* (Cambridge: Cambridge University Press).
Slote, Michael (1983) *Goods and Virtues* (Oxford: Oxford University Press).
Smith, Holly M. (1988) "Making Moral Decisions," *Nous*, vol. 22, pp. 89-108.
Taylor, Richard (1985) *Ethics, Faith, and Reason* (Englewood Cliffs, NJ: Prentice-Hall).
Thomas, Lawrence (1989) *Living Morally: A Psychology of Moral Character* (Philadelphia: Temple University Press).
Trianosky, Gregory (1990) "Natural Affection and Responsibility for Character: A Critique of Kantian Views of the Virtues," in Owen Flanagan and Amelie Rorty (eds.) *Identity, Character, and Morality* (Cambridge: MIT Press).
Trianosky, Gregory (1988) "Rightly Ordered Appetites: How to Live Morally and Live Well," *American Philosophical Quarterly*, vol. 25, pp. 1-12.
Trianosky, Gregory (1987) "Virtue, Action, and the Good Life: Toward a Theory of the Virtues," *Pacific Philosophical Quarterly*, vol. 68, pp. 124-47.
Von Wright, G. H. (1963) *The Varieties of Goodness* (London: Humanities Press).
Wallace, James (1978) *Virtues and Vices* (Ithaca: Cornell University Press).
Warnock, Geoffrey (1971) *The Object of Morality* (London: Methuen Press).
Williams, Bernard (1981) "Persons, Character, and Morality," in *Moral Luck* (Cambridge: Cambridge University Press).

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 27, Number 4, October 1990

THE RANGE OF OPTIONS

Michael J. Zimmerman

WHAT options we have in a situation can clearly be of great moral significance. I shall argue that our options are far more restricted than is commonly thought. If I am right, then my argument has important moral implications.

An option is an action that one can perform.[1] The sort of "can" at issue is that which J. L. Austin called "all-in."[2] Its analysis is controversial.[3] I won't try to analyze it here; I hope that what I have in mind will be clear enough if I say that it is the sort of "can" that lies at the heart of the debates on free will and determinism and on whether "ought" implies "can."

I shall proceed as follows. In section I, I shall present my argument, which will contain one premise (with three clauses) and a conclusion. In Section II I shall defend the principle of inference used. In Sections III-IV I shall defend the three clauses of the premise as best I can when they are taken individually. In Section VI, I shall defend the premise as a whole (by this time it will have undergone some modification). I shall conclude in Section VII with some brief observations about the moral implications of the argument.

I. THE ARGUMENT

Here is an initial, incomplete rendering of the argument:

(A) (1) In any situation, very many of one's apparent options are actions which are such that

(i) one cannot perform them unless one has certain conscious thoughts concerning them and

(ii) one does not in fact have these thoughts.

Hence:

(2) In any situation, very many of one's apparent options are actions which one cannot in fact perform.

Before moving to an evaluation of the argument, let me note just what is at stake in its conclusion. By "an apparent option" I don't mean an option which it in fact appears to someone (oneself or another) that one has; I mean, roughly, an option which a normal person would admit to having when asked whether it is (or was) an option that he or she has (or had) in the situation and when answering sincerely. For example, suppose that Joe, a normal person, forgot to keep an appointment. While it didn't occur to him to keep it, nonetheless, if he were asked whether it was an option of his, he would say that it was; indeed, he'd probably say that it was one he should have taken. Or again, suppose that Joe, eager to press his point of view, kept interrupting his interlocutor. While it didn't occur to him to keep his mouth shut, he would, if asked, admit that he could and should have done so. Or finally, suppose that Joe, listening to his favorite song on the radio, ran a red light and caused an accident. He'd admit that he could and should have been paying more attention to what he was doing. My contention (put bluntly and a little too boldly) is that, in all of these (and other such) cases, Joe is mistaken. None of the apparent options were actual options of his, and this is simply because performing them never crossed his mind.

Let us now turn to an evaluation of the argument. It may appear that it goes awry in a familiar way. Consider this argument, which, apart from the absence of quantifiers, seems analogous to mine:

(B) (1) (i) Jones cannot walk unless he moves his legs, and

(ii) Jones does not in fact move his legs.

Hence:

(2) Jones cannot in fact walk.

Clearly, something has gone wrong here, and the diagnosis is not hard to come by. Using obvious symbolism—obvious but, as we shall soon see, not entirely innocent—we may put the diagnosis as follows. It may at first be tempting to construe (B) as having this form (form "a"):

(Ba) (1) (i) $-m \rightarrow -P(w)$ &
(ii) $-m$
Hence:
(2) $-P(w)$

But while this form of argument is valid, it is apparent that (Ba1) does not capture what is meant by (B1). On the contrary, (B) would appear to have the following form:

(Bb) (1) (i) $-P(w \,\&\, -m)$ &
(ii) -m
Hence:
(2) $-P(w)$

But now, it will be pointed out, although the first premise is true, the argument is invalid. (It is assumed that personal possibility—that which is expressed by "P" and, in English, by "(all-in) can"—is wholly analogous to strict logical possibility.)

I think that this dismissal of (B) is quite correct. But now consider this argument:

(C) (1) (i) Smith cannot walk unless she has legs, and
(ii) Smith in fact has no legs.
Hence:
(2) Smith cannot in fact walk.

This seems much better than (B). Why? Presumably because it would *not* be a mistake to see it as having the following valid form:

(Ca) (1) (i) $-h \rightarrow -P(w)$ &
(ii) $-h$
Hence:
(2) $-P(w)$

But now we need to enquire: why is it that (B1i) and (C1i) should be seen to have such different logical forms when, on the surface, they appear so similar?

To answer this question, consider one way to modify (Bb) so that (B)'s form is apparently valid. The alteration is in the second clause of the premise:

(Bc) (1) (i) $-P(w \,\&\, -m)$ &
(ii) $-P(m)$
Hence:
(2) $-P(w)$

If Jones *cannot* move his legs, *then* he cannot walk. (The principle of inference presupposed here seems to be this:

(P1) $-P(p \,\&\, -q), -P(q) \vdash -P(p)$

Again, an analogy is apparently presupposed between personal possibility and strict logical possibility. For, where the latter is substituted for the former in (P1), we have an acceptable principle.[4]) This alteration to the argument, though, while rendering it valid, does not succeed in rendering it sound. For (Bc1ii), we may assume, is false; even if Jones doesn't move his legs, he can.

In this respect, though, arguments (B) and (C) differ. Not only *does* Smith have no legs, it is not personally *possible* for her to have any; that is, roughly, she can do nothing about her legless condition.[5] It is this implicit understanding of the difference between Jones's and Smith's situations that accounts for the difference between the arguments. That is, while (Ca) does capture the form of (C), the following would render explicit what is at issue in (C):

(Cd) (1) (i) $-P(w \,\&\, -h)$ &
(ii) $-h$ &
(iii) $-h \rightarrow -P(h)$
Hence:
(2) $-P(w)$

Given principle (P1), (Cd1) *yields* (Ca1). The proponent of (C) is implicitly relying on the truth of (Cd1), even though what is explicitly stated is captured by (Ca1).

Now, how does all this apply to my original argument (A)? Not surprisingly, I want to say that that argument is properly seen to have a form analogous to that of argument (C), not (B), and thus that it is to be understood along the lines of (Cd). Roughly, then, and with quantifiers omitted, what we have is an argument of this form:

(Ad) (1) (i) $-P(a \,\&\, -c)$ &
(ii) $-c$ &
(iii) $-c \rightarrow -P(c)$
Hence:
(2) $-P(a)$

More fully (although this isn't the final version):

(A') (1) In any situation, very many of one's apparent options are actions which are such that
(i) one cannot both perform them and not have certain conscious thoughts concerning them,
(ii) one does not in fact have these thoughts, and
(iii) if one does not have these thoughts, one cannot have them.
Hence:
(2) In any situation, very many of one's apparent options are actions which one cannot in fact perform.

What we need now to do is look more closely at the modal principle of inference presupposed and at the argument's premise.

II. The Principle of Inference

I have said that the modal principle of inference presupposed in (Bc)—and hence, by extension, in (Cd), (Ad), and finally (A')—seems to be the following:

(P1) $-P(p \,\&\, -q), -P(q) \vdash -P(p)$.

It is natural to treat (P1) as akin to familiar principles of the logic of strict logical necessity and possibility. If we do this, we must interpret 'p' and 'q' as propositional variables, and as a propositional connective, and 'P' as a propositional operator. (P1) may then be thought to be equally well rendered as

(P2) $N(p \rightarrow q), P(p) \vdash P(q)$

(where "N" expresses personal necessity or unavoidability), and this itself may be seen to be equivalent to

(P3) $N(p \rightarrow q), N(p) \vdash N(q)$.[6]

Each of these principles is extremely plausible, but how *exactly* are phrases of the form "$P(p)$" and '$N(p)$' to be rendered in English?

The best rendition, I think, is this: "$P(p)$" means the same as "it is personally possible for agent S at time t that p," and "$N(p)$" means the same as "it is personally necessary for S at t that p." (Just what *these* mean, I'll get to in a moment.) But here we run into a snag. So understood, (P3) (to which, it has just been said, (P1) is equivalent) has been used in a well-known recent argument for the thesis that free will is incompatible with determinism.[7] While that argument relies on several assumptions other than the assumption that (P3) is valid, nonetheless some sympathizers with compatibilism have claimed that it is (P3) that is to be rejected. Three objections have been raised: that (P3) is invalid when 'p' expresses a proposition having to do with the past[8]; that, in its use of "$N(p)$," (P3) presupposes what is impossible, *viz.*, that an agent can enter into a causal relation with a proposition[9]; and that (P3) begs the question against compatibilism[10]. I shall now address these concerns.

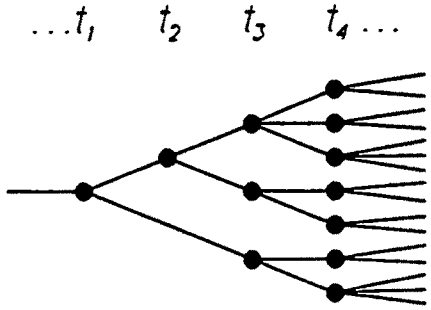
While I believe the first objection to be mistaken, the best way to deal with it for present purposes is simply to circumvent it by adopting a principle that is narrower than (P3) but which still serves to validate the conclusion of my argument. Thus I propose the following, where t is no later than either t' or t^*

> If it is not personally possible for S at t that event e occur at t' and event f not occur at t^*, and if it is not personally possible for S at t that f occur at t^*, it follows that it is not personally possible for S at t that e occur at t'.

That is, in symbolism more explicit than that used so far:

(P4) $-P_{S,t}(e_{t'} \,\&\, f_{t^*}), -P_{S,t}(f_{t^*}) \vdash -P_{S,t}(e_{t'})$.[11]

The second objection can be met by giving a definition of personal possibility that employs the all-in "can" in such a way as to dispel any fears of presupposing an impossible relation between agents and propositions. This definition is based on the following considerations. Life is a series of choices.[12] At each point of choice we are faced with a number of different directions in which we may travel. By travelling in one direction rather than another, we actualize one possible future and close off others. Each possible future is a segment of some possible world. Thus, at each point of choice, we have the option of actualizing one possible world rather than another.[13] The picture is roughly this:



Two comments are in order. First, this picture is inaccurate because each point of choice may have many more lines leading from it than are indicated here. (Nonetheless, the point of this paper is that the number of these lines is far smaller than is commonly believed.) Second, the picture conforms with (P4). If there is no line leading from a point of choice at t on which e occurs at t' and f does not occur at t^*, and if there is no line leading from that same point of choice on which f occurs at t^*, then there is no line leading from that point of choice on which e occurs at t'.

I now suggest the following recursive definition (where a possible world is understood as a logically

possible proposition that entails every proposition or its negation, and where "can" is all-in):

(D1) It is personally possible for S at t that p =*df* for some possible world W:
(i) W entails that p, and either
(ii) S exists at t and W is actual, or
(iii) for some choice c and time t':
(a) S can at t make c at t', and
(b) if (1) S were to make c at t' and
(2) whatever S cannot at t prevent from occurring at t' were to occur at t'
then either (3) W would obtain
or (4) it would be personally possible for S at t' that W obtain.[14]

The third objection is more difficult to handle, because what constitutes question-begging is itself very unclear. I believe that there is good reason to reject the objection simply on the basis that the assumption that (P4) is valid does not by itself entail that compatibilism is false; as noted earlier, other assumptions are required. Nonetheless, some may complain that (P4) *does* by itself pose a threat to compatibilism, in that many compatibilists presuppose a conditional analysis of "can" and (P4) implies that such an analysis is false. Thus these compatibilists, at least, won't find (P4) innocuous.

Now, whether (P4) is inconsistent with a conditional analysis of "can" of course depends on just what the analysis is. (P4) certainly *is* inconsistent with the following popular analysis:

(D2) S can do a =*df* if S were to choose to do a, S would do a.[15]

But this analysis is demonstrably false anyway.[16] Still, it must be acknowledged that perhaps some conditional, compatibilist analysis of "can" could be devised, the overturning of which would require appealing to (P4) (or some principle implied by (P4)). In this case, there would be a stand-off. Independent considerations would have to be summoned in favor of the analysis on the one hand and of (P4) on the other. I see every reason to think that the plausibility of (P4) would outweigh any consideration in favor of the analysis.[17] In saying this, I am presupposing what is surely true, namely, that the truth of compatibilism does not require the adequacy of any such analysis. Thus, while (P4) by itself may pose a threat to some compatibilists, I still believe that it poses no threat to compatibilism as such.

III. The Premise: First Clause

If we are to consider the first clause of the premise in isolation, we must consider this proposition:

(i) In any situation, very many of one's apparent options are actions which are such that it is not personally possible for one both that one perform them and that one not have have certain conscious thoughts concerning them.

Just what the "certain conscious thoughts" are will vary from case to case, but the general point is simply that very many actions are such that one cannot perform them unless one pays some attention to what one is doing; success in one's endeavors very frequently requires that one concentrate on the task at hand.

There may appear to be three broad classes of actions that do not satisfy this general condition. The first is that of unintentional actions. In reaching out for a second cup of coffee, I may "succeed" in knocking over the coffee pot. Doing this never crossed my mind; it certainly wasn't something that I was attending to. But here we may observe that, even though I was not attending to knocking over the coffee pot, I was attending to reaching out for a second cup of coffee; moreover, I would not have done the former unless I had done the latter, and doing the latter *required* my attention. In general, even unintentional action involves intentional action[18] and must therefore involve whatever conscious thoughts intentional action must involve.

The second class of apparent exceptions is that of routine or habitual actions. To what extent such actions can be performed unthinkingly is unclear. It surely is true that such actions can be performed without being the *focus* of one's attention, but whether or not they can be performed without the agent lending some *minimal* attention to them is controversial.[19] Let us assume, however, that some routine actions require not even minimal consciousness of them for their successful undertaking. (Perhaps tying one's shoelaces, scratching an itch, changing gears, and so on, are frequently actions of this sort.) At most this assumption requires admitting that many of one's apparent options are actions which are such that it *is* personally possible for one both that one perform them and that one not have certain conscious thoughts concerning them. It does *not* require that we reject (i).[20]

The third class[21] is that of actions whose initiation appears to require certain conscious thoughts but whose continuation does not. Presumably, in order to begin sunbathing I must pay some attention to what I am doing; but it seems that I can continue doing this quite mindlessly (perhaps I am engrossed in the novel I'm reading). Two points are in order here. First, just as with the second class of exceptions, this class does not imply that (i) is false. Second, even if *continuing* certain actions requires no thought, *desisting* from their continuation seems to require some (while my sunbathing may stop without my thinking of anything, it would appear that *I* cannot stop it without thinking about what I am doing), and this observation supports (i).

IV. The Premise: Second Clause

The proposition that we must now attend to is this:

(ii) In any situation, very many of one's apparent options are actions concerning which one does not in fact have conscious thoughts.

There are two types of evidence for this claim. The first is personal. For my own part, when I concentrate on what I am doing, I do not attend to what I may otherwise do. When I do not concentrate on what I am doing, then usually my mind wanders; again, I do not attend to my options. I strongly suspect that most people are similar to me in these respects, although I have no firm evidence to support this. Of course, it can and does at times happen that I take the time to reconnoiter the territory, to canvass my options. Even here, though, I attend to few options at a time. And at this point the second type of evidence— experimental evidence—is pertinent. Apparently humans are capable of attending to at most five to seven independent chunks of information at once.[22] Now, regardless of just how this claim is to be interpreted, it is evident that consciousness is highly selective. No one is capable of attending to many independent options at once; hence no one does.

5. The Premise: Third Clause

The proposition to be considered here is this:

(iii) In any situation, very many of one's apparent options are actions which are such that, if one does not in fact have conscious thoughts concerning them, it is not personally possible for one that one have such thoughts.

This will most likely appear to be the most vulnerable part of my argument. Why should it be necessary to *have* the thoughts in question in order for it to be *possible* for one to have them? Indeed, this claim is indefensible as it stands, but I think that a strong case can be made for a somewhat diluted version of it.

First, we should ask how in general it might be personally possible for someone at t that he or she have a certain conscious thought at t'. There seem to be just two main possibilities. Either one has a choice at t as to whether one has the thought at t', or one does not. If one has a choice, then one can ensure its occurrence in one of two main ways. Either one already has the thought at t (or at some time t^* intermediate between t and t') and can maintain it through t' (i.e., so act that one contributes to its sustenance through t'); or one does not already have the thought at t (or t^*)—or one has it at t (or t^*) but fails to maintain it through t'—but can so act that one contributes to its occurrence at t'. If, on the other hand, one does not have a choice at t as to whether one has the thought at t', it is still personally possible for one at t that one have the thought at t' if, but only if, it is personally necessary for one at t that one have it then; that is, if, but only if, the thought occurs anyway at t', independently of any choice that one makes at t (or t^*). In summary, then: one can have a thought just in case either one can maintain it, or one can contribute to it, or it will occur anyway.

Let us now turn to our immediate options. By "an immediate option," I don't mean an action that can be completed immediately. There are no such actions, since all actions, even the most basic, take time to complete. Nonetheless, if there are any actions we can perform at all, there are actions that we can *begin* immediately to complete, and it is such actions that I call our immediate options.

The point here is this. Given (i), very many of my immediate options are such that I cannot (begin to) perform them without having certain conscious thoughts. Thus, if it is personally possible for me now immediately to perform them, it must be personally possible for me now immediately to have the thoughts in question. Given the ways in which having such thoughts is personally possible, having them immediately is personally possible for me now only if I can now maintain them, or I can now contribute to them, or they will occur anyway. But it is being assumed (in the antecedent of (iii)) that I do

not now have these thoughts; hence I cannot now maintain them. Moreover, I wish to argue, I cannot now contribute to their *immediate* occurrence; for this would require that the contribution be a basic action, and yet contributing to the having of a conscious thought is never (or hardly ever[23]) a basic action. That is, one can accomplish such contribution only by way of some intermediate action which itself contributes to the having of the thought; this intermediate action must take place at a time intermediate between now and the time of the thought; hence the thought cannot be immediately produced. Thus we are left only with this possibility: having the thought immediately is personally possible for me now only if the thought occurs independently of any choice of mine now. Now, while it may be very likely that *some* thought (or thoughts) will immediately occur to me (and hence very likely that it is personally possible for me now to have *some* thought), every thought, I contend, (with one possible exception) is such that *it* is very unlikely immediately to occur to me; hence (with one possible exception) every thought is such that it is very unlikely that it is personally possible for me now immediately to have it.[24]

Why should this be, though, and what is the possible exception? Consider, first, that the number of independent conscious thoughts any one of which one is in principle capable of having at one time is extremely large (call it N); and second, that the maximum number of independent conscious thoughts all of which one is capable of having at one time is very small (five to seven, as mentioned before—call it n). Thus, if all thoughts were equiprobable, the likelihood that on some occasion a particular one of them should occur would be n/N—i.e., very low indeed. Now, of course, we cannot assume that all thoughts are equiprobable; some are far more likely to occur than others, given the agent's past history, propensities, circumstances, etc. For example, one is much more likely to think of what to have for dinner than to wonder how much wood a woodchuck could really chuck. Indeed—and this is the possible exception—it may be that having a certain thought at t' is personally necessary for someone at t *because of some choice* made by him or her *prior* to t. In such a case it may even be *likely* (from a perspective prior to t'), rather than unlikely, that the thought in question will occur at t'. But even if this is so for some thoughts, it will not be so for many others. Indeed, it *cannot* be; for, given that at most n independent thoughts can occur at once, the increased likelihood of one thought tends to render others *less* likely. Thus, ironically, *whatever exceptions there are prove*[25] *the general rule.* This is so even if we ignore "irrelevant" thoughts altogether and concentrate only on those that are of potential concern to the agent; for the number of such thoughts is still very large (call it N^*) and hence n/N^* is still very small.[26]

We may thus conclude that, in any situation, very many of one's apparent *immediate* options are actions which are such that, if one does not in fact have conscious thoughts concerning them, it is *very likely* not personally possible for one that one have such thoughts. This is significant because, conjoined with the other clauses of the premise, it yields (with a caveat to be explained in the next section) the conclusion that the number of lines immediately issuing from a point of choice is far smaller than one would normally think. And, of course, this restriction on the number of immediate options entails a restriction on the number of non-immediate, or remote, options. Still, it could yet be that the number of remote options is very large, and so let me now address this matter directly.

Given what was said above, it is personally possible for me now that I have certain conscious thoughts in the non-immediate future—thoughts which I do not now have—only if I contribute to them or they occur anyway. That any particular such thought should occur independently of any choice or action of mine is just as unlikely in this case as in the case of my immediately having the thought. But, in this case, as opposed to the former case, there is no reason in principle to think that I cannot contribute to such thoughts; for time to do this is in principle available. Nonetheless, there is still good reason to think that, for almost any particular such thought, it is highly unlikely that I can contribute to it. Let me explain.

I can contribute to the having of a thought in one of two ways: *intentionally* or *unintentionally*. I suggest that one can *intentionally* contribute to the having of a conscious thought only if, at the time at which the contribution is initiated, one *has* the thought. Since we are working under the assumption that I do *not* presently have the thought, the only contribution to the later having of the thought that I can now initiate is unintentional. Now, it seems to

me that most of my conscious thoughts are indeed brought about in just this way: I do something which unintentionally results in my having a certain thought. But how likely is it that one particular thought rather than another would[27] be *unintentionally* brought about on some particular occasion—*especially* when the number of immediate options is as small as I have argued? Again, it seems that, given the very large pool of candidate thoughts (even when restricted to "relevant" thoughts) and the very small number of thoughts that can be had at any one moment, we must conclude that the likelihood is, in almost all cases, very small.[28]

Here an objection may be made. It may seem that there is an intermediate path between simply unintentionally producing a thought and intentionally producing it, and that is the strategy of intentionally scanning my apparent options and thereby, given sufficient time and skill in scanning, latching consciously on to almost all such options, thus rendering them genuine options after all. While none of the thoughts that I would thereby produce would itself be intentionally produced, I would intentionally bring it about that such thoughts occur to me. Doesn't the fact that such scanning is frequently possible show that the likelihood of my having conscious thoughts concerning particular apparent options is in fact often quite *high*?

The answer, I think, is still no. Although I concede that having the leisure to scan one's apparent options will raise the probability of each such option's being thought of and hence becoming a real option, two important points must be borne in mind. First, if undertaking the scanning is *itself* to be possible for one, one must have a conscious thought concerning *it*—and this itself may well not be very likely. Second, it is not enough to latch consciously on to an apparent option for it to become a real option; one must latch on to it *at the right time*. If at t_1, during the course of scanning, I think of a certain possible course of action at t_3 but at t_2, during the course of further scanning, I relinquish the thought and it doesn't recur to me at t_3, then at t_3 the option that requires the thought will remain merely apparent. Now, I concede that having a thought raises considerably the probability of its occurring later[29], and so my having the thought at t_1 will raise the probability of its recurring at or during t_3. Indeed, it *may* even render it likely that I will have the thought (perhaps it will occur to me to maintain the thought from t_1 through t_3; or perhaps it will occur to me to make a reminder to myself that I'll be likely to consult in time). But, as before, this ironically *supports* my general thesis. For if a certain thought is rendered likely, then, given the cap on the number of thoughts that may occur simultaneously, *other* thoughts are rendered less likely, so that, *in general* (even if there are certain particular exceptions), the likelihood of a certain "relevant" thought occurring to me *at a certain time* will be low, whether or not scanning has taken place.[30]

I therefore conclude that the following modification to (iii) is true:

> (iii') In any situation, very many of one's apparent options are actions which are such that, if one does not in fact have conscious thoughts concerning them, it is *very likely* not personally possible for one that one have such thoughts.

VI. The Argument: Final Version

The final version of my argument must obviously reflect the modification just made to the third clause of the premise. But getting the final version of the premise is not just a matter of conjoining (i), (ii), and (iii'), because (and this is the caveat mentioned in the last section) it may be that very many x are F, very many x are G, and very many x are H, while it is not the case that very many x are F and G and H. I believe, however, that in the present case this fact presents no problem. For the considerations that were advanced separately in support of (i), (ii), and (iii') taken individually may, I believe, be advanced jointly in support of a "merged" version of these propositions. That is, it seems reasonable to believe that very many of the very many apparent options that satisfy (i) also satisfy (ii) and (iii'), i.e., that there is a high degree of overlap between them. If this seems too complacent or contentious, consider the following. Suppose that we made the conservative estimate that the "very many" of one's apparent options at issue in each of (i), (ii), and (iii') constituted only 90% of those options, and that we also made the assumption that there was the *minimal* possible overlap between the options which satisfied (i), (ii), and (iii'). The result would still be that 70% of one's apparent options satisfied all of (i), (ii), and (iii')—and this still constitutes "very many." Hence I claim that the considerations brought to bear in the

preceding three sections furnish strong support for the following argument:

(A") (1) In any situation, very many of one's apparent options are actions which are such that
(i) it is not personally possible for one both that one perform them and that one not have certain conscious thoughts concerning them,
(ii) one does not in fact have these thoughts, and
(iii) if one does not have these thoughts, it is very likely not personally possible for one that one have them.

Hence:

(2) In any situation, very many of one's apparent options are actions which are such that it is very likely not in fact personally possible for one that one perform them.

VII. Moral Implications

Joe forgot to keep his appointment; he kept interrupting his interlocutor; he ran a red light. If my argument is sound, it is likely that he couldn't have done otherwise. If one ought to do only what one can do, then it is likely not the case that Joe ought to have done otherwise. If one is responsible only for what one could have avoided, then it is likely not the case that Joe is responsible for what he did. Both theses are plausible; if they are correct, then my argument implies that there is likely much less wrong done and much less responsibility incurred than is commonly believed. Even if the theses are not correct, some close modification of them probably is, and so my argument probably still has this implication.[31] At the very least, if (as is surely plausible) what one ought to do and what one is responsible for doing are in part functions of what options one has or had, then my argument implies that obligation and responsibility are often in an important respect "subjective," tied to the agent's state of mind.

At this point it should be acknowledged that there is an ambiguity to "can," even when it is circumscribed as at the outset of this paper. First, there is the question whether a person under compulsion can, in the relevant sense, resist the compulsion. Can one, at gunpoint, defy a gunman's order? A "liberal" on this issue will say yes (for all it takes is the choice, admittedly a difficult one but not an impossible one, to decline to cooperate), while a "conservative" will say no.[32] Then there is the question whether a person who cannot intentionally do something can nevertheless sometimes do it. Can one open a safe even when one has no idea what the combination is?[33] Again, a "liberal" will say yes (for all it takes is a series of finger movements, easy to perform) while a "conservative" will say no.[34]

Unlike some, I am willing to declare myself a liberal. That is, *as long as one has the requisite conscious thoughts*, one can, I believe, defy the gunman and one can open the safe. This has been the position implicitly adopted in this paper, for my argument applies to options even when so liberally construed. But what of the two moral theses just mentioned? Do they concern the liberal "can" or some more conservative "can"? An answer to this question need not be given here, for if my argument is sound with respect to the liberal "can" it is *ipso facto* sound with respect to some more conservative "can" and hence will still have the moral implications cited.[35]

Throughout this paper I have construed an option to be an *action* that one can perform. But what of *omissions*? It may seem that there is a large class of morally significant behavior that is left untouched by the foregoing observations, even if they are otherwise accurate: that of satisfying negative duties. I can, it may be said, satisfy the duties not to kill, not to commit adultery, not to bear false witness, and so on, quite mindlessly; indeed, I do so whenever I am absorbed in routine activities.

The easiest response to this claim is capitulation. We should, I am sure, grant that omissions in general and the satisfaction of negative duties in particular are sometimes options that we have, and I would be quite content if my case for the restricted nature of our options were itself to be restricted to our "positive" options (genuine actions that we can perform) so that our "negative" options ("not-doings" that are personally possible for us) are left untouched.[36]

Another response, however, is this. If one ought to do only what one can avoid doing, then the satisfaction of negative duties does not constitute an exception to my argument after all. Even if omissions are not actions[37] and even if they are sometimes options that we have, if "ought" implies "can avoid" it will not be the case that I ought not to kill, or commit adultery, or bear false witness unless I can do these things. If, given my argument, I can rarely do these things, then I rarely have the duties in question.

Finally, it may be claimed that my argument could be used as a basis for the thesis that we ought to

ensure (or try to ensure) that we have certain conscious thoughts—those thoughts without which the duties that we would otherwise have would become defunct. This seems plausible to me, but whether or not it is correct can be determined only by appeal to a general theory of obligation, something that I shall not undertake here. Still, this much can be said: if my argument is sound, and if "ought" implies "can," it is very likely that one has the obligation to (try to) ensure that one has certain conscious thoughts only if it occurs to one to (try to) ensure this.[38]

University or North Carolina at Greensboro

Received January 26, 1990

NOTES

1. More precisely: an option is either an action or an omission that one can accomplish. On omissions see section VII below.

2. [1] p.319.

3. Some have said that one can "all-in" do something just in case one has the ability and the opportunity to do it ([20] p. 325); others have denied this ([14] p. 242). Many have touted a conditional analysis (including [2] p. 282, [10] p. 159, [12] p. 104, [18] pp. 90-91, [21] p. 57); others have rejected this ([4] p. 345, [14] pp. 248-50, [24] p. 114ff.)—more on this later.

4. In the "weakest" modal system, *T*, the following is an axiom (rather than a principle of inference):

$$L(p \rightarrow q) \rightarrow (Lp \rightarrow Lq).$$

(See [11] p. 31, where "*L*" and "*M*" are used to express strict logical necessity and possibility, respectively.) This is provably equivalent to:

$$(\text{-}M(p \ \& \ \text{-}q) \ \& \ \text{-}Mq) \rightarrow \text{-}Mp.$$

5. See (D1) in section II for a more precise account.

6. Again, compare the logic of strict logical necessity and possibility. See note 4 above.

7. [24] p. 94. Actually, "*N*" is not agent- and time-bound in the argument as it is implicitly in (P3).

8. [22] p. 19.

9. [19] p. 84. Cf. [24] p. 67.

10. [9] p. 71. Cf. [24] pp. 18, 102-4; [8] pp. 432-40.

11. Why opt for something so complicated? Why not simply accept the following?

> If *S* cannot at t do action *a* at *t'* without doing action *b* at *t**, and if *S* cannot at *t* do *b* at t*, it follows that *S* cannot at *t* do *a* at *t'*.

The answer is this. Even if this principle is valid (as I believe), it is inapplicable to my argument (specifically to Aé1i) because having conscious thoughts is not itself an action (although it is perhaps something that one "does," in some broad sense).

12. In saying this I don't assume that all choices must be conscious. Nonetheless I am inclined to believe that this is so. See note 20 below.

13. Cf. [6] p.16ff.

14. Clause (ii) accommodates those relatively rare occasions where *S* is already travelling in some direction but cannot ever again make a choice.

15. See the philosophers mentioned in note 3 above as ones who have advocated a conditional analysis. All of these writers have embraced something close to (D2). That (P4) is inconsistent with (D2) is shown by the following illustration. Suppose that Smith is a sensible person. He knows that he cannot make a cake from scratch without using flour, and so he would choose to make a cake from scratch only if flour were available. Moreover, if he did choose to make a cake from scratch, he'd succeed in doing so. Of course, he would not succeed in making a cake from scratch without flour if, for some bizarre reason, he were to choose to do that. It is also true, though, that he has no flour and that, unbeknownst to him, he has no way to get any; thus, if he were to choose to get flour, he'd fail to do so. Let *S* be Smith, m be the proposition that Smith makes a cake from scratch, and f be the proposition that Smith has flour. What this case tells us, given (D2), is this (temporal subscripts omitted): $\text{-}P_S(m \ \& \ \text{-}f)$, $\text{-}P_S(f)$, but $P_S(m)$. Thus we have a counterexample to (P4).

16. See [24] p. 119. Cf [7], p. 17ff.

17. Cf. [24] p. 122.

18. Or something close to it. See [26] p. 140.

19. On the question of the focus of attention, see [25] Ch. 10, section 2; [13] section 35.

20. I believe that the assumption is in fact much less far-reaching than this alone would suggest, however. Routine or habitual actions (given that they are wholly unthinking) seem to be performed on "automatic pilot," to use a common and suggestive analogy. Such actions are typically intentional and may involve decisions, but these decisions are *not* choices between presently competing alternatives, for the course has already been set. Routine or habit *dictates* the decisions, so that other options, even if they were once genuine, are no longer so; they have been ruled out. Only when one consciously backs off from the routine (because some object or event has forced itself upon one's consciousness—cf. [25] p. 163) do other options emerge as genuine candidates for choice. Thus I would contend that a conscious appraisal of one's situation is in fact necessary if one is to have genuine alternatives. Cf. [16] pp. 241-43, [17] p. 316.

21. Pointed out to me by Terry McConnell.

22. [16] pp. 236, 246ff.

23. It may be that some people (e.g., skilled mental arithmeticians) can sometimes immediately conjure up some thoughts. But if this proves an exception to my argument, I take it to be one that is negligible. Surely it is far, far more common that thoughts can be and are evoked only nonbasically.

24. The phrase "(un)likely that it is personally possible" may be disconcerting, involving as it does a double modality. The best antidote I can think of is simply to keep in mind that it is the *all-in* "can" that is at issue. It doesn't sound odd to say, "He can probably do that."

25. In the sense of "confirm," and not of "test."

26. Sample "relevant" thoughts include thoughts concerning: this project; that project; what to have for dinner; what's upsetting Joe; what to get Jane for her birthday; paying the car insurance premium; committee work... It would be tedious to extend this list; it could, and in almost all cases would, be very long.

27. Or could. See clause (iiib4) of (D1).

28. It is important that I do *not* presently have the thought in question and hence that any contribution to its later occurrence is unintentional. I concede that my presently having the thought would raise considerably the probability of its occurring as a result of some action that I can now perform.

29. See the last note.

30. One particular sort of exception to this general claim should be explicitly noted, however, and that is where the time in question is sufficiently *extended*. Compare my calling Sue *at* noon and my calling her *by* noon. It is much more likely that I can do the latter than the former, and this is because what constitutes the "right" time for the relevant thought to occur is much longer in the latter case.

31. I believe that "ought" implies "can." I do not believe that responsibility implies avoidability, although I do think that one cannot be responsible for doing something unless one *believed* one could avoid doing it (as long as one is not responsible for not believing this). See [27] p. 22 and Ch.4, section 10.

32. A representative liberal: Thorp ([23] pp. 8-9). A representative conservative: Dennett ([5] p. 133).

33. See [24] p. 230, n.9.

34. A representative liberal: Feldman ([6] pp. 24-25). A representative conservative: Lemos ([15] p. 302).

35. In [27] pp. 85-6 I argued that one typically lacks control, in the second of the conservative senses just cited, over one's thoughts and that this would afford an excuse for much thoughtless behavior. My contention in this paper is that one typically lacks control even in the liberal senses over one's thoughts.

36. Still, it should be noted that *intentional* omissions—refrainings—will often require conscious thoughts, just as actions often do. Thus my argument would apply to them.

37. Although they sometimes involve them. See [26] pp. 183-84.

38. My thanks to Josh Hoffman, John King, Terry McConnell, Al Mele, and Gary Rosenkrantz for helpful comments on earlier drafts.

BIBLIOGRAPHY

[1] Austin, J.L. "Ifs and Cans." In [3], pp. 295-322.
[2] Ayer, A.J. *Philosophical Essays* (London: Macmillan, 1963).
[3] Berofsky, Bernard, ed. *Free Will and Determinism* (New York: Harper and Row, 1966).
[4] Chisholm, Roderick M. "J.L. Austin's Philosophical Papers." In [3], pp. 339-345.
[5] Dennett, Daniel C. *Elbow Room* (Cambridge: MIT Press, 1984).
[6] Feldman, Fred. *Doing the Best We Can* (Dordrecht: D. Reidel, 1986.)
[7] Fischer, John Martin. "Introduction: Responsibility and Freedom." In *Moral Responsibility*, edited by John Martin Fischer (Ithaca: Cornell University Press, 1986), pp. 9-61.
[8] Flint, Thomas P. "Compatibilism and the Argument from Unavoidability." *Journal of Philosophy*, vol. 84 (1987), pp. 423-40.
[9] Foley, Richard. "Compatibilism and Control over the Past." *Analysis*, vol. 39 (1979), pp. 70-4.
[10] Hobbes, Thomas. *Leviathan* (London: Collier-Macmillan, 1962).
[11] Hughes, G.E. and Cresswell, M.J. *An Introduction to Modal Logic* (London: Methuen, 1968).
[12] Hume, David. *An Inquiry concerning Human Understanding* (New York: Bobbs-Merrill, 1955).
[13] Husserl, Edmund. *Ideas* (London: Collier-Macmillan, 1962).
[14] Lehrer, Keith. "'Can' in Theory and Practice: A Possible Worlds Analysis." In *Action Theory*, edited by Myles Brand and Douglas Walton (Dordrecht: D. Reidel, 1976), pp. 241-70.
[15] Lemos, Ramon M. "Duty and Ignorance." *Southern Journal of Philosophy*, vol. 18 (1980), pp. 301-12.
[16] Mandler, George. "Consciousness: Respectable, Useful, and Probably Necessary." In *Information Processing and Cognition*, edited by Robert L. Solso (Hillsdale: Lawrence Erlbaum, 1975), pp. 229-54.
[17] Mandler, G. and Kessen, W. "The Appearance of Free Will." In *Philosophy of Psychology*, edited by S. C. Brown (London: Macmillan, 1974), pp. 305-24.
[18] Moore, G.E. *Ethics* (Oxford: Oxford University Press, 1978).
[19] Narveson, Jan. "Compatibilism Defended." *Philosophical Studies*, vol. 32 (1977), pp. 83-7.
[20] Nowell-Smith, P.H. "Ifs and Cans." In [3], pp. 322-39.
[21] Schlick, Moritz. "When Is a Man Responsible?" In [3], pp. 54-63.
[22] Slote, Michael. "Selective Necessity and the Free-will Problem." *Journal of Philosophy*, vol. 79 (1982), pp. 5-24.
[23] Thorp, John. *Free Will.* (London: Routledge and Kegan Paul, 1980).
[24] van Inwagen, Peter. *An Essay on Free Will* (Oxford: Clarendon Press, 1983).
[25] Vernon, M.D. *The Psychology of Perception* (Harmondsworth: Penguin, 1962).
[26] Zimmerman, Michael J. *An Essay on Human Action* (New York: Peter Lang, 1984).
[27] Zimmerman, Michael J. *An Essay on Moral Responsibility* (Totowa: Rowman and Littlefield, 1988).

AMERICAN PHILOSOPHICAL QUARTERLY
Volume 27, Number 4, October 1990

The Editor's Page

Argumentation and Rhetoric in Philosophical Method

There are two very different modes of writing philosophy. The one pivots on inferential expressions such as "because," "since," "therefore," "has the consequence that," "and so cannot," "must accordingly," and the like. The other bristles with adjectives of approbation or derogation—"evident," "sensible," "untenable," "absurd," "inappropriate," "unscientific," and comparable adverbs like "obviously," "foolishly," etc. The former relies primarily on inference and argumentation to substantiate its claims, the latter primarily on the rhetoric of persuasion. The one seeks to secure the reader's (or auditor's) assent by reasoning, the other by an appeal to values.

Consider the following passage from Nietzsche's *Genealogy of Morals* (with characterizations of approbation/derogation indicated by being italicized):

> It is in the sphere of contracts and legal obligations that the moral universe of guilt, conscience, and duty, (*"sacred" duty*) took its inception. Those beginnings were *liberally sprinkled with blood*, as are the beginnings of *everything great on earth*. (And may we not say that ethics has never lost its *reek of blood* and torture—not even in Kant, whose categorical imperative *smacks of cruelty*?) It was then that the *sinister knitting together* of the two ideas guilt and pain first occurred, which by now have become quite inextricable. Let us ask once more: in what sense could pain constitute repayment of a debt? In the sense that to make someone suffer was *a supreme pleasure*. To behold suffering gives pleasure, but to cause another to suffer affords an *even greater pleasure*. This *severe statement* expresses an old, powerful, *human, all too human sentiment*—though the monkeys too might endorse it, for it is reported that they heralded and preluded man in the devising of *bizarre cruelties*. There is no feast without cruelty, as man's entire history attests. Punishment, too, has its *festive features*. (Friedrich Nietzsche, *The Geneology of Morals*, Essay II, Sect. 6.)

Not only is the passage replete with devices of evaluative (i.e. positive/negative) characterization, but observe too the total absence of inferential expressions. We are, clearly, *invited* to draw certain unstated evaluative conclusions. But the inference "Man is by *nature* given to cruelty, and so cruelty—being a natural and congenial tendency of ours—is not something bad, something deserving condemnation" is left wholly implicit: It is hinted at but never stated. In consequence, reason can gain no fulcrum for pressing the plausible objection: "And why should something natural automatically be therefore good: why should the primitiveness of a sentiment or mode of behavior safeguard it against a negative evaluation?" By leaving the reader to his own conclusion-drawing devices, Nietzche relieves himself of the labor of argumentation. Not troubling to formulate his position, he feels no need to give it *support*; he is quite content to *insinuate* it.

By contrast to the preceding Nietzsche passage, consider the following ideologically kindred passage from Hume's *Treatise* (with evaluative terms italicized and inferential terms capitalized):

> Now, SINCE the distinguishing impressions by which moral good or evil is known are nothing but particular pains or pleasures, IT FOLLOWS that in all inquiries concerning these moral distinctions IT WILL BE SUFFICIENT TO SHOW the principles which make us feel a satisfaction or uneasiness from the survey of any character, IN ORDER TO SATISFY US WHY the character is *laudable* or *blamable*. An action, or sentiment, or character, is *virtuous* or *vicious*; WHY? BECAUSE its view causes a pleasure or uneasiness of a particular kind. In giving a reason, THEREFORE, for the pleasure or uneasiness, we sufficiently explain the vice or virtue. To have the sense of virtue is nothing but to feel a satisfaction of a particular kind from the contemplation of a character. The very feeling constitutes our praise or admiration. We go no further; nor do we inquire into the cause of the satisfaction. WE DO NOT INFER a character to

> be *virtuous* BECAUSE it pleases; but in feeling that it pleases after such a particular manner we in effect feel that it is *virtuous*. The case is the same as in our judgments concerning all kinds of beauty, and tastes, and sensations. Our approbation is IMPLIED in the immediate pleasure they convey to us. (David Hume, *Treatise of Human Nature*, Bk. III, Pt. I, Sect. 2.)

While for Nietzche cruelty is effectively a virtue because people are held to be statistically pleased by engaging in its practice, for Hume it is something negative only in that people are held to be statistically displeased by witnessing it. The positions differ but their ideological kinship is clear: both writers agree that cruelty is not something that is inherently bad as such.

What is also clear, however, is that these kindred positions are advanced in very different ways. In the Nietzche passage, the "argumentation ratio" of inferential to evaluative expressions is 0:12, in the Hume passage it is 9:6. Hume, in effect, seeks to *reason* his readers into agreement; Nietzche to *coax* them into it.

Reflection on the contrast between the argumentative and the rhetorical modes of philosophical exposition leads to the realization that these two styles are congenial to somewhat different objectives. The demonstrative/argumentative (inferential) mode is efficient for securing a reader's assent to certain claims, to influencing one's *beliefs*. The rhetorical (evocative) mode is optimal for inducing a reader to adopt certain preferences, to shaping or influencing one's *priorities and evaluations*.

The argumentative, *apodictic* (or probative) mode of philosophical exposition is by nature geared to enlisting the reader's assent to certain theses or theories. It is coordinated to a view of philosophy that sees the discipline in *information-oriented* terms, as preoccupied with the answering of certain questions: the solution of certain cognitive problems. It aims primarily to *convince* by way of reasoning.

By contrast, the rhetorical, *prohairetic* (or evocative) mode of philosophical exposition is by nature geared to securing acceptance with respect to *evaluations*: to enlisting the reader's agreement to certain priorities or appraisals. It is preoccupied with evaluation, with forming—or reforming—our sensibilities with respect to the *value* and, above all, the *importance* of various items. It is bound up with a view of philosophy that sees the discipline in *axiological* terms, as an enterprise that has as its prime task the securing of certain evaluative determinations and the establishment of certain prizings and priorities. It aims primarily to *induce* people to an evaluative standpoint.

To exert rational pressure on a reader's values without using arguments that are themselves already value-invoking, one must deploy (or reshape) this person's body of experiences. Here too providing information can help—but only by way of influencing the sensibility—the reader's way of looking at things. Accordingly, it is here that the rhetorical method comes into its own by enabling an exposition appear to—and if need be modify—a reader's body of experiences in order to induce a suitable adjustment of evaluations. There are, of course, many ways to realize this sort of objective. A survey of suitably related case studies, a survey of selected historical episodes ("History teaching by examples"), or a vividly articulated fiction can orient a reader's evaluative sentiments in a preplanned direction, particularly when supplemented by a suitably tendentious interpretative exegesis. The assembling of illustrative episodes—real or constructed—can render good service in this regard as the philosophical methodology of Ludwig Wittgenstein amply illustrates. And so, of course, can pure invective, if sufficiently clever in its articulation.

Two distinct views of the mission of the enterprise being at issue with the demonstrative and evaluative approaches to philosophy, any debate over the respective merits of the two modes of philosophical exposition is thereby inextricable from a dispute about the nature of philosophy. The quarrel is ultimately one of ownership: to whom does the discipline of philosophy properly belong, to the argumentative demonstrators or to the evaluative charmers?

This contest over the ownership of philosophy has been going on since the very inception of the subject. Among the Presocratics, the Milesians founded a "nature philosophy" addressed primarily at issues we shall nowadays classify as explanatorily scientific, while such thinkers as Xenophanes, Heraclitus, and Pythagoras took an evaluative approach to philosophy, illustrated by the following dictum of the last-named:

> Life is like a festival; just as some come to the festival to compete, some to ply their trade, but the best people come as spectators, so in life slavish men go hunting for fame or gain, the philosophers for the truth. (Frag. 278, Kirk & Raven.)

In 19th Century German philosophy, Hegel and his school typified the scientific/demonstrative approach, while the "post-moderns" who were their opponents—Schopenhauer, Kierkegaard, Nietzsche—all exemplify the axiological/rhetorical approach. In the 20th century, the scientistic movement represented by logical positivism vociferously insisted on using the methodology of demonstration, while their anti-rationalistic opponents among the existentialists and also among the neo-Romantic theoreticians of Spain (preeminently including Unamuno and Ortega y Gasset) resorted extensively to predominantly literary devices to promulgate their views—to such an extent that their demonstration-minded opponents sought to exile their work from philosophy into literature, journalism, or some such less "serious" mode of intellectual endeavor.

In this connection we see as clearly as anywhere else the tendency among philosophers towards defining the entire subject in such a way that their own work is central to the enterprise and that their own favored methodology becomes definitive for the way in which work in the field should properly be done. The absence of that urbanity which enables one to see other people's ways of doing things as appropriate and acceptable is unfortunately the most widespread and characteristic failing of practicing philosophers. For the fact is that while individual *philosophers* generally have no alternative but to choose one particular mode of philosophizing as forms of their allegiance, *philosophy* as such has to accommodate both. Philosophy as such is broader than any one philosopher's philosophy.

The irony is that philosophers simply cannot dispense altogether with the methodology they affect to reject and despise.

Even the most demonstration-minded philosopher cannot avoid entanglement in evaluation by rhetorical devices. For one cannot argue for everything, "all the way down," so to speak. At some point one must invite assent through mere rhetoric.

On the other hand, even the most evaluation-minded philosopher cannot altogether avert argumentation. For a reliance on certain *standards* of assessment is inescapably present in those proffered evaluations, and this issue of appropriateness cannot be addressed satisfactorily without some recourse to reasons.

Ironically then, the two modes of philosophy are locked into an uneasy but indissoluble union. Neither side can feel altogether comfortable about using the methodology favored by the other, and yet neither side can manage altogether to get on without it.

BOOKS RECEIVED

Azar, Larry, *Twentieth Century in Crisis* (Dubuque: Kendall/Hunt Publishing Company, 1990), 317 pp., $24.95.

Balaban, Oded, *Subject and Consciousness* (Savage: Rowman & Littlefield Publishers, 1990), 240 pp., $32.50.

Benjamin, Andrew, ed., *The Lyotard Reader* (Oxford and Cambridge: Basil Blackwell Ltd., 1989), 425 pp.

Buck-Morss, Susan, *The Dialectics of Seeing* (Cambridge: The MIT Press, 1990), 493 pp., $29.95.

Callahan, Daniel, *What Kind of Life* (New York: Simon and Schuster, 1990), 318 pp., $19.95.

Dahlhaus, Carl, *The Idea of Absolute Music*, Roger Lustig, trans. (Chicago: University of Chicago Press, 1990), 176 pp., $29.95.

Daniels, Norman, *Am I My Parents Keeper?* (New York: Oxford University Press, 1990), 193 pp., $9.95.

Devine, Philip E., *The Ethics of Homicide* (Notre Dame: University of Notre Dame Press, 1990), 264 pp., $10.95.

Edel, Abraham, *Interpreting Education* (New York: Prometheus Books, 1990), 340 pp., $18.95.

Fox, Richard M. and DeMarco, Joseph P., *Moral Reasoning* (Fort Worth: Holt, Rinehart and Winston, Inc., 1990), 341 pp., $18.00.

Griffiths, Paul J., *et al.*, *The Realm of Awakening* (New York and Oxford: Oxford University Press, 1989), 399 pp., $49.95.

Hollier, Denis, *Against Architecture*, Betsy Wing, trans. (Cambridge: The MIT Press, 1990), 201 pp., $19.95.

Jamme, Christoph and Poggeler, Otto, eds., *Phanomenologie im Widerstreit: zum 50*, Todestag Edmund Husserls (Frankfurt am Main: Euhrkamp, 1989), 372 pp.

Janich, Peter, *Euklinds Erbe: Ist der Raum dreidimensionol?* (Munich: Beck, 1989), 246 pp.

Kemp, T. Peter and Rasmussen, David, *The Narrative Path* (Cambridge: The MIT Press, 1990), 121 pp., $10.95.

Kitcher, Philip and Salmon, Wesley C., eds., *Scientific Explanation* (Minneapolis: University of Minnesota Press, 1989), 528 pp.

Kline, George L., ed., *Alfred North Whitehead* (Maryland: University Press of America, 1989), 206 pp.

Leslie, John, ed., *Physical Cosmology and Philosophy* (New York: Macmillan Publishing Company, 1990), 277 pp., $39.50.

Levy, Ze'ev, *Baruch or Benedict* (New York: Peter Lang Publishing, 1989), 224 pp.

Long, A. A. and Sedley, D. N., *The Hellenistic Philosophers* (New York: Cambridge University Press, 1989), 512 pp., $24.95.

LoSurdo, Domenico, *Hegel Und Das Deutsche Erbe* (Dahl-Rugenstein Verlag), 531 pp.

Marion, Jean Lue, *Reduction et donation: Recherches sur Husserl Heidegger et la phenomenologie* (Paris: Presses Universitaires de France, 1989), 312 pp.

Mittelstrass, Jürgen, *Der Flug der Eule: Von der Vernunft der Wissenschaft und der Aufgabe der Philosophie* (Frankfurt am Main: Suhrkamp, 1989), 333 pp.

Nerlich, Graham, *Values and Valuing* (Oxford: Clarendon Press, 1989), 217 pp., $45.00.

Ozmon, Howard A. and Craver, Samuel M., *Philosophical Foundations of Education, Fourth Edition* (Columbus: Merrill Publishing Co., 1990), 400 pp.

Paquet, Leonce, *et al.*, *Les Presocratiques: Bibliographie Analytique (1879-1980)* (Montreal: Les Editions Bellarmin, 1989).

Parsons, Keith M., *God: And The Burden of Proof* (Buffalo: Prometheus Books, 1989), 156 pp., $34.95.

Rapp, Friedrich and Wiehl, Reiner, eds., *Whitehead's Metaphysics of Creativity* (Albany: State University of New York Press, 1990), 223 pp., $49.50 (Paper, $16.95).

Segura, Armando, *Emmanuel: Principia Philosophica (Madrid: Encuentro, 1982), 503 pp.*

Segura, Armando, *Principios de Filosofia de la Historia* (Madrid: Encuentro, 1985), 162 pp.

Smith, Gary, ed., *Benjamin* (Chicago: University of Chicago Press, 1990), 263 pp., $32.95.

Volney, Constantin-Francois, *Oeuvres I and II (2 vols.)* (Paris: Fayard, 1989), 684 pp. (vol. I), 504 pp. (vol. II).

Warning, Rainer, ed., *Estetica de la Recepcion*, Ricardo Sanchez and Ortiz deUrbina, trans. (Madrid: La Balsa de la Medusa, 1979), 313 pp.

Waterfield, Robin, *Before Eureka* (New York: St. Martin's Press, 1990), 124 pp., $29.95.

Wren, Thomas E., *The Moral Domain* (Cambridge: The MIT Press, 1990), 414 pp., $35.00.

Zarka, Yves Charles and Bernhardt, Jean, *Thomas Hobbes* (Paris: Presses Universitaires De France, 1990), 415 pp.

CORRIGENDA

Volume 26 (October 1989)

Howard McGary, "Forgiveness," pp. 343-51. The following endnotes were omitted on page 351:

14. Laurence H. Davis, *A Theory of Action* (Englewood Cliffs: Prentice-Hall, 1979), p. 15.

15. Robert Kraut, "Feelings in Context," *The Journal of Philosophy*, vol. 83 (1986), pp. 647-48.

16. C. G. Montefiore and H. Lowe, *A Rabbinic Anthology* (New York: Schocken Books, 1974), Chap. 19, and G. F. Moore, *Judaism*, Vol. 1, Chapter 6 and Vol. 2, Chapter 5 (Cambridge, MA: Harvard University Press, 1927-1930).

17. I am grateful to Mary Gibson, Douglas Husak, Brian McLauglin, Uma Narayan, Laurence Thomas, and members of the philosophy department at the University of Illinois at Chicago for their comments and suggestions to earlier drafts of this paper. Of course, I bear responsibility for any errors that remain.

Volume 26, July 1989

Robert Almeder, "Scientific Realism and Explanation," pp. 173-185. The following endnotes were omitted on page 185:

19. "Blind Realism," *Erkenntnis*, vol. 26 (1987), pp. 1-8, and "Fallibilism, Coherence and Realism," *Synthese*, 68 (1986), pp. 213-23. The argument offered in the latter essay is the same one offered in the former; the former essay repeats the argument in the later because a full statement of the general thesis in one place required as much.

20. In the third section of "Blind Realism" I offered arguments for the view that we do not in fact have, and will never have, a reliable decision procedure for picking out, establishing or otherwise determining *which* statements succeed in correctly describing the external world. Also, for another argument urging that realism does not require acceptance of the correspondence theory of truth, see Brian Ellis, "What Science Aims to Do," in *Images of Science*, pp. 54 ff.

21. See, for example, Rorty's reply to Hilary Putnam and Richard Boyd in *Philosophy and the Mirror of Nature* (Princeton: Princeton University Press, 1979), p. 284. Rorty responds to Putnam and Boyd's argument that realism explains the success of theories by seeming to claim that scientific success needs no explanation. He says: "The fact that new theories often go wrong just where the old ones say they might is not something that requires an explanation." Also, Fine attacks realism because he sees it as based upon the false belief that science, all science, is successful. For Fine, because it is simply false that all science is successful, we do not need realism to explain the success of science ("The Natural Ontological Attitude," p. 104, n. 8). Fine's objection, however, is a strawman of the realist's position, which does not seek to explain the success of *all* scientific theories rather than those that have been, or are, predictively successful in the long run.

22. For a full discussion on William Craig and F. P. Ramsey's attempt to eliminate all reference to theoretical entities, and why the procedure is bound to fail, see Carl Hempel, "The Theoretician's Dilemma" in *Minnesota Studies in the Philosophy of Science*, vol. 2 (Minneapolis, Minnesota: University of Minnesota Press, 1958); and Isreal Scheffler, *The Anatomy of Inquiry* (New York: Knopf, 1963), pp. 193 ff.

23. See my "Blind Realism." I would like to thank Paul Humphreys, Nicholas Rescher, Clifford Hooker, and an anonymous referee for their comments on an earlier version of this paper. I am also grateful to the Hambidge Center and the Center for the Philosophy of Science at the University of Pittsburgh for providing the protected solitude necessary for the completion of this paper.

PHILOSOPHY DOCUMENTATION CENTER

Mailing Lists

RENTAL AGREEMENT

All mailing lists are available for one-time rental use. Placement of orders with the Philosophy Documentation Center constitutes the renter's agreement not to 1) reuse the lists, 2) photocopy the lists for reuse, 3) allow anyone else to use the lists or part thereof, and 4) place the lists into a database for any purpose whatsoever.

STANDARD LISTS — U. S. & Canada

Philosophers	Count*	Cheshire Price
Entire List	10,700	$ 195
United States only	9,550	175
Canada only	1,150	25
Philosophy Departments		
All philosophy departments	2,190	$ 65
PhD major	130	25
MA major or PhD minor	100	25
BA major or MA minor	695	35
BA minor or AA major	290	30
Other, e.g., courses in philosophy with no degree specialization	975	40
Philosophers by Specialty		
Aesthetics	1,280	$ 45
American Philosophy	290	30
Ancient Philosophy	640	35
Contemporary Philosophy	915	40
Epistemology	1,044	40
Ethics	2,390	65
History of Modern Philosophy	625	35
Logic	1,085	45
Medieval Philosophy	290	30
Metaphysics	1,485	50
Oriental Philosophy	290	30
Phenomenology and Existentialism	800	40
Philosophy of Education	144	25
Philosophy of History	915	40
Philosophy of Language	525	35
Philosophy of Law	300	30
Philosophy of Religion	1,485	55
Philosophy of Science	1,095	45
Social and Political Philosophy	1,258	45

*Counts are approximated

CUSTOMIZED LISTS

Custom lists of philosophers or philosophy departments can be provided for a base charge of $25 plus $20 per 1,000 labels. In addition to the above specialties, you can request mailing labels of philosophers who are interested in any area of philosophy *or* you can request labels of all philosophers teaching in PhD and/or MA departments. Almost any combination is possible. For example, labels for philosophers in PhD departments specializing in ethics can be provided.

GUARANTEE — 100% Deliverable

The Philosophy Documentation Center will refund $.50 for each piece that was undeliverable due to an incorrect name or address. We must receive undeliverable pieces within 60 days after the labels are shipped to you.

GUARANTEE — Lowest Cost

The PDC guarantees that its prices for mailing lists are the lowest available. You can rent the List of U.S. and Canadian Philosophers for less than $20 per 1,000. Our other lists are also competitively priced.

GEOGRAPHICAL SELECTIONS — Free

All lists are generated in ZIP code order. You can request labels by state or province (up to ten states or provinces). There is no extra charge for this service.

CODING OF LISTS — Free

All lists can be run with a ten character code, which is printed on the right side of line one. There is no charge for this service.

GRADUATE ASSISTANTS — Optional

The Standard Lists of Philosophers include approximately 1,200 graduate assistants. You may exclude them from your labels *or* you may choose graduate assistants only.

COUNTS

Computer searches of the database to obtain counts for customized lists cost $25 each. This computer time charge is waived if you purchase a set of labels.

TYPES OF LABELS

Four-up Cheshire labels are standard and are included in the prices that are given. Add 10% to the prices for three-up self-adhesive labels.

TIME REQUIRED

Orders are usually shipped the next day. However, we guarantee that your order will be shipped within five working days.

SHIPPING & HANDLING

Orders within the U.S. are shipped U.P.S. at a cost of $5 per list. The charge for special mailing services such as overnight delivery is $20. Orders shipped outside the U.S. are sent via the best available carrier.
Shipping and handling charges will be waived if orders are prepaid and do not require special mailing services.

COPYRIGHT

WRITE OR CALL

To order labels or for further information, please contact Mrs. Cindy Richards, Philosophy Documentation Center, Bowling Green State University, Bowling Green, Ohio 43403-0189 U.S.A.; (800) 444-2419, *or* outside the U.S. (419) 372-2419, FAX: (419) 372-6987.

PUBLIC AFFAIRS QUARTERLY

The PUBLIC AFFAIRS QUARTERLY, a scholarly journal devoted to current issues in social and political philosophy, commenced publication in January, 1987 with the cooperation of the Philosophy Documentation Center. The journal specializes in contributions which examine, in the light of philosophical reflections and assessments, matters that figure on the current agenda of public policy, building upon the contemporary interest in tightly focused philosophical case studies of particular issues in such areas as social and economic justice, public welfare; individual entitlements, rights and duties; inheritance, taxation and distributive justice in general; population policy, abortion, euthanasia; environmental problems, science policy; the social and political status of women, senior citizens, minorities and other social groups; arms control, war and deterrence; loyalty, duty and patriotism; ethical issues in medicine, business and the professions; criminality, criminal justice and punishment; and similar topics.

The PUBLIC AFFAIRS QUARTERLY seeks to enhance the quality of our understanding of issues of public policy by publishing essays that bring philosophical depth and sophistication to the consideration of matters on the agenda of public debate that would otherwise be left to the tender mercies of political rhetoric and journalistic oversimplification. However, insofar as clarification can be separated from advocacy, the journal stands on the side of elucidation rather than partisanship.

Subscriptions to the journal should be sent to the Philosophy Documentation Center, Bowling Green State University, Bowling Green, Ohio 43403-0189 ($26 for individuals; $72 for institutions).

Submissions for possible publication and other correspondence should be addressed to:

The Editor
Public Affairs Quarterly
Department of Philosophy
University of Pittsburgh
Pittsburgh, PA 15260

HISTORY OF PHILOSOPHY QUARTERLY

The HISTORY OF PHILOSOPHY QUARTERLY is a scholarly journal which commenced publication in January 1984 with the cooperation of the Philosophy Documentation Center. The whole of each issue is devoted entirely to articles. (There are no book reviews and no discussion notes, but general surveys on particular topics may be published from time to time.)

The HPQ interests itself particularly in papers that cultivate philosophical history in the spirit of *philosophia perennis*. Ideally, its contributions will regard work in the history of philosophy and in philosophy itself as parts of a seamless whole, treating the work of past philosophers not only in terms of historical inquiry, but also as a means of dealing with issues of ongoing philosophical concern. The journal favors that approach to philosophical history, increasingly prominent in recent years, which refuses to see the boundary between philosophy and its history as an impassible barrier, but regards historical studies as a way of dealing with problems of continued interest and importance.

Submissions along these lines will be welcomed. They may be as short as 3,000 words and as long as 8,000. The editor also plans to commission occasional reviews of current work in particularly active areas of investigation.

Subscriptions to the journal should be sent to the Philosophy Documentation Center, Bowling Green State University, Bowling Green, Ohio 43403-0189 ($29 for individuals; $98 for institutions).

Submissions for possible publication and other correspondence should be addressed to:

The Editor
History of Philosophy Quarterly
Department of Philosophy
University of Pittsburgh
Pittsburgh, PA 15260